(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application:	November 19, 2001
Application Number:	Japanese Patent Application No. 353640/2001
Applicant(s):	Hitachi, Ltd.

February 22, 2002

Commissioner,
Patent Office Kozo Oikawa (seal)

Certificate No. 2002-3009734

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００１年１１月１９日

出願番号
Application Number: 特願２００１－３５３６４０

[ST.10/C]: [ＪＰ２００１－３５３６４０]

出願人
Applicant(s): 株式会社日立製作所

２００２年 ２月２２日

特許庁長官
Commissioner,
Japan Patent Office 及川耕造

出証番号 出証特２００２－３００９７３４

【書類名】　特許願

【整理番号】　H101741

【提出日】　平成13年11月19日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F　7/00

【発明者】

【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

【氏名】　安田　知弘

【発明者】

【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

【氏名】　西川　哲夫

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社　日立製作所

【代理人】

【識別番号】　100091096

【弁理士】

【氏名又は名称】　平木　祐輔

【手数料の表示】

【予納台帳番号】　015244

【納付金額】　21,000円

【その他】　国等の委託研究の成果に係る特許出願（平成１３年度新エネルギー・産業技術総合開発機構（再）委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　　要約書　1

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　核酸塩基配列のアセンブル方法及びアセンブル装置

【特許請求の範囲】

【請求項１】　第１の核酸塩基配列上で固定長のウィンドウを移動させながら、当該ウィンドウによって切り取られた配列と一致する部分配列を末端領域に有する第２の核酸塩基配列を探索する工程と、

前記工程で探索された第２の核酸塩基配列と前記第１の核酸塩基配列とがアセンブル可能かどうか判定する工程と、

前記工程でアセンブル可能と判定された場合、前記第１の核酸塩基配列と前記第２の核酸塩基配列とをアセンブルする工程とを含むことを特徴とする核酸塩基配列のアセンブル方法。

【請求項２】　第１の核酸塩基配列上で固定長のウィンドウを移動させながら、当該ウィンドウによって切り取られた配列と一致する部分配列を末端領域に有する第２の核酸塩基配列を探索する工程と、

前記工程で探索された第２の核酸塩基配列と前記第１の核酸塩基配列とがアセンブル可能かどうか判定する工程と、

前記工程でアセンブル可能と判定された場合、前記第１の核酸塩基配列と前記第２の核酸塩基配列とをアセンブルする工程とを含み、

前記工程でアセンブルされた核酸塩基配列を改めて第１の核酸塩基配列として前述の工程を反復することを特徴とする核酸塩基配列のアセンブル方法。

【請求項３】　複数の核酸塩基配列について、各核酸塩基配列を識別する情報と当該核酸塩基配列の末端領域に位置する固定塩基長の部分配列とを関連づけてテーブルに登録する工程と、

第1の核酸塩基配列に基づき最初のコンセンサス配列を構築する工程と、

前記テーブルを参照して、前記コンセンサス配列の一部と一致する部分配列を有する核酸塩基配列を探索する工程と、

前記工程で探索された核酸塩基配列と前記コンセンサス配列間の前記部分配列に連続する配列同士を比較し、探索された核酸塩基配列が前記コンセンサス配列にアセンブル可能かどうかを判定する工程と、

前記工程でアセンブル可能と判定された場合、前記コンセンサス配列に前記核酸塩基配列をアセンブルしてコンセンサス配列を再構築する工程とを有することを特徴とする核酸塩基配列のアセンブル方法。

【請求項４】　請求項３記載の核酸塩基配列のアセンブル方法において、前記第1の核酸塩基配列として未処理の核酸塩基配列中で最長の塩基長を有する配列を選択することを特徴とする核酸塩基配列のアセンブル方法。

【請求項５】　複数の核酸塩基配列を配列長の降順にソートする第１の工程と、

前記複数の核酸塩基配列について、各核酸塩基配列を識別する情報と当該核酸塩基配列の末端領域に位置する固定塩基長の部分配列とを関連づけてテーブルに登録する第２の工程と、

未処理の複数の核酸塩基配列の中から配列の長さが最長の核酸塩基配列を１つ選び最初のコンセンサス配列を構築する第３の工程と、

前記コンセンサス配列上で固定長ウィンドウを移動させつつ、前記テーブルを参照して、前記固定長ウィンドウによって切り取られた配列と一致する部分配列を有する未処理の核酸塩基配列を探索する第４の工程と、

前記コンセンサス配列と前記第４の工程で探索された未処理の核酸塩基配列とを比較し、両者がアセンブル可能であるか否かを判定する第５の工程と、

前記第５の工程でアセンブル可能と判定された場合、前記コンセンサス配列に前記第４の工程で探索された核酸塩基配列をアセンブルし、コンセンサス配列を再構築する第６の工程とを含み、

前記固定長ウィンドウが前記コンセンサス配列の全領域を走査するまで第４の工程から第６の工程を反復し、さらに未処理の核酸塩基配列が残っていれば前記第３の工程から前記第６の工程を反復することを特徴とする核酸塩基配列のアセンブル方法。

【請求項６】　請求項３～５のいずれか１項記載の核酸塩基配列のアセンブル方法において、１つの核酸塩基配列に対して前記テーブルに登録する前記固定塩基長の部分配列の数を指定する工程を含むことを特徴とする核酸塩基配列のアセンブル方法。

【請求項７】　請求項３～６のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記テーブルに登録する前記固定塩基長の部分配列を抽出する前記核酸塩基配列の末端領域の範囲を指定する工程を含むことを特徴とする核酸塩基配列のアセンブル方法。

【請求項８】　請求項３～７のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記テーブルに登録する前記固定塩基長の部分配列の塩基長を10以上32以下とすることを特徴とする核酸塩基配列のアセンブル方法。

【請求項９】　請求項３～７のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記テーブルを一回参照したとき検出され、かつ前記コンセンサス配列とアセンブル不可能であると判定されるエントリの数の期待値の上限値cを指定する工程を含み、

前記複数の核酸塩基配列の数をN、各核酸塩基配列から選択される固定塩基長の部分配列の数をKとするとき、次式（１）を満足する整数sを前記テーブルに登録する固定塩基長の部分配列の塩基長とすることを特徴とする核酸塩基配列のアセンブル方法。

$$s \geq \frac{1}{2}\log\frac{KN}{c} \qquad \cdots(1)$$

【請求項１０】　請求項３～９のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記コンセンサス配列を格納するデータ構造に双方向リストを使用することを特徴とする核酸塩基配列のアセンブル方法。

【請求項１１】　請求項３～１０のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記固定塩基長の部分配列を該固定塩基長部分配列の長さと無関係な固定数の計算機ワードで表現することを特徴とする核酸塩基配列のアセンブル方法。

【請求項１２】　請求項３～１１のいずれか１項記載の核酸塩基配列のアセンブル方法において、前記テーブル中、予め指定された回数以下の出現回数をもつキーに対応するエントリだけを利用することを特徴とする核酸塩基配列のアセンブル方法。

【請求項１３】　入力核酸塩基配列の末端領域に設定される固定塩基長の部

分配列に関するパラメータを入力する入力手段と、

複数の入力核酸塩基配列に対して、各核酸塩基配列を識別する情報と前記入力手段によって入力されたパラメータに従って当該核酸塩基配列から抽出した固定塩基長の部分配列とを関連づけてテーブルに登録する手段と、

前記テーブルを参照してコンセンサス配列の一部と一致する部分配列を有する核酸塩基配列を探索する手段と、

前記手段で探索された核酸塩基配列と前記コンセンサス配列とを比較し、両者がアセンブル可能かどうかを判定する判定手段と、

前記判定手段でアセンブル可能と判定された場合、前記コンセンサス配列と前記探索された核酸塩基配列とをアセンブルしてコンセンサス配列を再構築する手段とを含むことを特徴とする核酸塩基配列アセンブル装置。

【請求項１４】　請求項１３記載の核酸塩基配列アセンブル装置において、前記入力手段は、入力核酸塩基配列上の前記固定塩基長の部分配列の位置を配列の末端からの関係でグラフィカルに表示する表示部を有し、ユーザが指定した固定長部分配列位置を表示に即座に反映させることを特徴とする核酸塩基配列アセンブル装置。

【請求項１５】　請求項１３又は１４記載の核酸塩基配列アセンブル装置において、前記入力手段は、１つの入力核酸塩基配列に対して抽出すべき前記固定塩基長の部分配列の数及び前記固定塩基長の部分配列の塩基数を入力することを特徴とする核酸塩基配列アセンブル装置。

【請求項１６】　請求項１３記載の核酸塩基配列アセンブル装置において、前記入力手段は、前記テーブルに登録する固定塩基長の部分配列の長さを、当該テーブルを参照したときに検出される偶然の一致の数の期待値によって指定することを特徴とする核酸塩基配列アセンブル装置。

【請求項１７】　請求項１３～１６のいずれか１項記載の核酸塩基配列アセンブル装置において、前記テーブルに登録された前記固定塩基長の部分配列の配列毎の出現頻度を表すグラフィカルな表示及び／又は数値を表示する表示部を有することを特徴とする核酸塩基配列アセンブル装置。

【請求項１８】　請求項１７記載の核酸塩基配列アセンブル装置において、

前記出現頻度の上限を指定する手段と、前記テーブルから前記出現頻度が指定された上限を超えるエントリを削除する手段とを備えることを特徴とする核酸塩基配列アセンブル装置。

【請求項１９】　請求項１３～１８のいずれか１項記載の核酸塩基配列アセンブル装置において、前記コンセンサス配列にアセンブルされた各入力核酸塩基配列を、前記コンセンサス配列の一部と一致する前記入力核酸塩基配列の固定塩基長の部分配列の位置とともに表示する手段を備えることを特徴とする核酸塩基配列アセンブル装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、多数の核酸塩基配列を高速にクラスタリング及びアセンブルする方法に関する。

【０００２】

【従来の技術】

国際共同プロジェクト及び米国ベンチャー企業により、2000年6月にヒトゲノムの塩基配列決定の完了が宣言された。4色蛍光色素やキャピラリを使用したDNAシーケンサの普及など、DNA配列決定技術の進歩に伴い、E.coli、S.cerevisiaeを始めとする数十種類の微生物や、C. elegans、D.melanogasterなど多細胞生物の全ゲノム配列が決定され、ヒトゲノムのドラフト配列も利用可能となった。他にもマウスやイネなど、様々な生物種のゲノムプロジェクトが進行中である。

【０００３】

ゲノム配列の解析が進む一方で、発現している遺伝子について調べるために、mRNAの解析が行なわれている。mRNAは、遺伝子が発現する際、ゲノムDNAから生成されるRNA分子で、遺伝子の機能発現の過程で不可欠な物質である。mRNA分子は分解しやすいが、逆転写により容易にmRNAよりも安定な物質であるcDNAに転換できるため、cDNAの形で解析されることが多い。cDNAをシングルパス配列解析して得られた配列は、ESTと呼ばれる。ESTには様々な利用価値があるが、そのひとつがmRNA配列を得ることである。

図１３は、mRNA由来のESTのクラスタリング及びアセンブル処理の概要を説明する図である。

【０００４】

mRNA 1301をcDNAに転換したとき、5'末端を含む完全長cDNAを得ることは困難であり、これらのcDNAに基づくEST 1302は、図１３のように、通常5'端が不揃いな配列である。細胞や組織の全RNAから作成したcDNAライブラリ由来のESTを解析する場合には、EST配列集合1303のみが得られ、各ESTがどのmRNAから得られたものか予め知ることができない。EST 1302を収集した配列集合1303を、配列間の類似部分1305に基づき各配列を結合（アセンブル）し、矢印1304で象徴的に示すように小さい集合に分割（クラスタリング）することで、同一mRNAから得られたESTを同定し、さらにmRNAの配列を部分的に再構成した配列1306を得ることができる。

【０００５】

ヒトの場合、mRNAはタンパク質数に対応し10万種類以上存在すると言われ、入力として与えられたESTなどの配列データをクラスタリング及びアセンブルすることで、これらのmRNA配列それぞれに対応するアセンブリが得られることが理想である。現在、米国公共機関のデータベースには、未整理のヒト由来ESTが約200万配列、遺伝子毎のクラスタにクラスタリングされたESTを含めたヒトmRNA由来の遺伝子の配列が約150万配列存在する。ゲノム配列決定が進み、遺伝子の機能解析へと研究の焦点が移りつつあり、解析が必要なmRNA由来の配列数も、さらに増加していくと推測される。

【０００６】

アセンブルの技術は、ゲノム配列決定にも必須である。ゲノム配列の決定には、主にショットガン法が用いられる。ショットガン法による配列決定では、長いDNAを多数の細かい断片に分解してクローニングを行ない、各断片の配列を決定し、配列アセンブルを行って全体の配列を決定する。例えば、大腸菌のゲノム配列は、約4639K塩基で、通常必要とされる冗長度10のショットガン法で決定するためには、DNAシーケンサによる一回の泳動で得られる配列の長さが500塩基程度であることを考慮すると、$4.639\times10^{6}\times10/500=9.278\times10^{5}$配列のアセンブルが

必要になる。C. elegans、マウス、ヒト等の高等生物では、ゲノムサイズがさらに2～3桁大きいため、これらのゲノム決定に要する配列数は1千万～1億に達すると推定される。今後さまざまな生物のゲノム配列が決定されていく中で、アセンブルの対象となる塩基配列の数は、さらに増加していくことが予想される。

【０００７】

【発明が解決しようとする課題】

莫大な数の核酸塩基配列について、各配列の相互の関係を調べ、クラスタリングやアセンブルを行うのは、計算時間の観点から困難である。配列のクラスタリング及びアセンブルでは、配列間に存在するオーバーラップをいかに効率的に探索するかが課題となる。単純に、あらゆる配列間でオーバーラップを探索すると、配列数の2乗のオーダーの組み合わせを探索する必要があり、配列数の増加に伴い処理時間の急激な増加を招く。しかし、クラスタリング及びアセンブルの処理全体のオーダーは、配列数の2乗のオーダーよりもずっと小さいオーダーであることが望ましい。

【０００８】

クラスタリング及びアセンブルのためのオーバーラップの探索を効率的に行う手法としては、Huang, X. and Madan, A., Genome Research,9:868-877, 1999の方法が挙げられる。しかし、依然として処理する必要があるオーバーラップの数は、配列数の2乗のオーダーとなり、クラスタリング及びアセンブル処理全体も配列数の2乗のオーダーとなってしまう。クラスタリング及びアセンブル処理の対象となる配列数は、これまで増加の一途をたどってきており、今後もさらに増加することが予想される。

本発明は、このような従来技術の問題点に鑑み、配列のクラスタリング及びアセンブル処理を入力配列数の2乗よりも小さいオーダーの計算量で行ない、多数の核酸塩基配列を高速にクラスタリング及びアセンブルする方法及び装置を提供することを目的とする。

【０００９】

【課題を解決するための手段】

本発明は、配列のクラスタリング及びアセンブルにおけるオーバーラップ探索

を効率よく行うために、以下の核酸塩基配列のアセンブル方法を提供する。

すなわち、本発明による核酸塩基配列のアセンブル方法は、第１の核酸塩基配列上で固定長のウインドウを移動させながら、当該ウインドウによって切り取られた配列と一致する部分配列を末端領域に有する第２の核酸塩基配列を探索する工程と、前記工程で探索された第２の核酸塩基配列と第１の核酸塩基配列とがアセンブル可能かどうか判定する工程と、前記工程でアセンブル可能と判定された場合、第１の核酸塩基配列と第２の核酸塩基配列とをアセンブルする工程とを含むことを特徴とする。

【００１０】

本発明による核酸塩基配列のアセンブル方法は、また、第１の核酸塩基配列上で固定長のウインドウを移動させながら、当該ウインドウによって切り取られた配列と一致する部分配列を末端領域に有する第２の核酸塩基配列を探索する工程と、前記工程で探索された第２の核酸塩基配列と第１の核酸塩基配列とがアセンブル可能かどうか判定する工程と、前記工程でアセンブル可能と判定された場合、第１の核酸塩基配列と第２の核酸塩基配列とをアセンブルする工程とを含み、前記工程でアセンブルされた核酸塩基配列を改めて第１の核酸塩基配列として前述の工程を反復することを特徴とする。

【００１１】

本発明による核酸塩基配列のアセンブル方法は、また、複数の核酸塩基配列について、各核酸塩基配列を識別する情報と当該核酸塩基配列の末端領域に位置する固定塩基長の部分配列とを関連づけてテーブルに登録する工程と、第1の配列に基づき最初のコンセンサス配列を構築する工程と、テーブルを参照して、コンセンサス配列の一部と一致する部分配列を有する核酸塩基配列を探索する工程と、前記工程で探索された核酸塩基配列とコンセンサス配列間の前記部分配列に連続する配列同士を比較し、探索された核酸塩基配列がコンセンサス配列にアセンブル可能かどうかを判定する工程と、前記工程でアセンブル可能と判定された場合、コンセンサス配列に当該核酸塩基配列をアセンブルしてコンセンサス配列を再構築する工程とを有することを特徴とする。

【００１２】

最初のコンセンサス配列を構築するための第1の配列としては、未処理の核酸塩基配列中で最長の塩基長を有する配列を選択する。１回のアセンブルが終了する度に、テーブルからコンセンサス配列にアセンブルされた核酸塩基配列に由来するエントリを削除するのが好ましい。

【００１３】

本発明による核酸塩基配列のアセンブル方法は、また、複数の核酸塩基配列を配列長の降順にソートする第１の工程と、複数の核酸塩基配列について、各核酸塩基配列を識別する情報と当該核酸塩基配列の末端領域に位置する固定塩基長の部分配列とを関連づけてテーブルに登録する第２の工程と、未処理の複数の核酸塩基配列の中から配列の長さが最長の核酸塩基配列を１つ選びコンセンサス配列を構築する第３の工程と、コンセンサス配列上で固定長ウィンドウを移動させつつ、前記テーブルを参照して、固定長ウィンドウによって切り取られた配列と一致する部分配列を有する未処理の核酸塩基配列を探索する第４の工程と、コンサンサス配列と第４の工程で探索された未処理の核酸塩基配列とを比較し、両者がアセンブル可能であるか否かを判定する第５の工程と、第５の工程でアセンブル可能と判定された場合、コンセンサス配列に第４の工程で探索された核酸塩基配列をアセンブルし、コンセンサス配列を再構築する第６の工程とを含み、固定長ウィンドウがコンセンサス配列の全領域を走査するまで第４の工程から第６の工程を反復し、さらに未処理の核酸塩基配列が残っていれば第３の工程から第６の工程を反復することを特徴とする。

【００１４】

前記方法は、１つの核酸塩基配列に対して前記テーブルに登録する固定塩基長の部分配列の数を指定する工程を含むことができる。

また、前記テーブルに登録する固定塩基長の部分配列を抽出する核酸塩基配列の末端領域の範囲を指定する工程を含むことができる。

【００１５】

前記テーブルに登録する固定塩基長の部分配列の塩基長は、処理速度低下を防ぐために、テーブル参照時に検出される配列間オーバーラップと無関係なエントリ数を抑制する必要があるため、少なくとも10塩基以上とするのが好ましく、か

つ、1計算機ワードが32bitsの計算機であれば2ワード、64bitsの計算機ならば1ワードで表現可能な32塩基以下とすることが好ましい。

【００１６】

前記方法は、前記テーブルを一回参照したとき探索され、かつコンセンサス配列とアセンブル不可能であると判定されるエントリの数の期待値の上限値cを指定する工程を含み、複数の核酸塩基配列の数をN、各核酸塩基配列から選択される固定塩基長の部分配列の数をKとするとき、後述する式（１）を満足する整数sを前記テーブルに登録する固定塩基長の部分配列の塩基長とすることが、さらに好ましい。

コンセンサス配列を格納するデータ構造には双方向リストを使用するのが好ましい。

【００１７】

また、固定塩基長の部分配列を該固定塩基長部分配列の長さと無関係な固定数の計算機ワードで表現するのが好ましい。クラスタリング結果は、計算機の主記憶装置に蓄積せず、クラスタが完成するごとに逐次出力するのが好ましい。

また、前記テーブル中、予め指定された回数以下の出現回数をもつ部分配列に対応するエントリだけを利用するのが好ましい。

【００１８】

本発明は、また、上記の方法によるアセンブル（クラスタリング）処理を実行するために必要とされる入力配列の選択、パラメータの入力、クラスタリング及びアセンブル処理の経過表示、及び結果表示を行うためのグラフィカルなユーザインタフェースを提供する。

【００１９】

本発明による核酸塩基配列アセンブル装置は、入力核酸塩基配列の末端領域に設定される固定塩基長の部分配列に関するパラメータを入力する入力手段と、複数の入力核酸塩基配列に対して、各核酸塩基配列を識別する情報と入力手段によって入力されたパラメータに従って当該核酸塩基配列から抽出した固定塩基長の部分配列とを関連づけてテーブルに登録する手段と、前記テーブルを参照してコンセンサス配列の一部と一致する部分配列を有する核酸塩基配列を探索する手段

と、前記手段で探索された核酸塩基配列とコンセンサス配列とを比較し、両者がアセンブル可能かどうかを判定する判定手段と、判定手段でアセンブル可能と判定された場合、コンセンサス配列と探索された核酸塩基配列とをアセンブルしてコンセンサス配列を再構築する手段とを含むことを特徴とする。

【００２０】

入力手段は、入力核酸塩基配列上の固定塩基長の部分配列の位置を配列の末端からの関係でグラフィカルに表示する表示部を有し、ユーザが指定した固定長部分配列位置を表示に即座に反映させるものとすることができる。

入力手段は、また、１つの入力核酸塩基配列に対して抽出すべき固定塩基長の部分配列の数及び固定塩基長の部分配列の塩基数を入力するものとすることができる。

入力手段は、前記テーブルに登録する固定塩基長の部分配列の長さを、当該テーブルを参照したときに検出される偶然の一致の数の期待値によって指定してもよい。

【００２１】

核酸塩基配列アセンブル装置は、前記テーブルに登録された固定塩基長の部分配列の配列毎の出現頻度を表すグラフィカルな表示及び／又は数値を表示する表示部を有するのが好ましい。更に、出現頻度の上限を指定する手段と、前記テーブルから出現頻度が指定された上限を超えるエントリを削除する手段とを備えることが好ましい。

また、コンセンサス配列にアセンブルされた各入力核酸塩基配列を、コンセンサス配列の一部と一致する入力核酸塩基配列の固定塩基長の部分配列の位置とともに表示する手段を備えるのが好ましい。

【００２２】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を説明する。

【００２３】

本発明のクラスタリング及びアセンブル方法では、配列間のオーバーラップがもつ性質に着目した。図３の例に示すように、２つの配列間のオーバーラップ部

303，306は、必ず片方の配列（図３では配列301及び配列304）の末尾部分と、片方の配列（図３では配列302及び配列304）の先頭部分を含む。そこで本発明では、図１に示したように、入力配列101の先頭及び末尾の長さsの部分配列102を、固定長部分配列テーブル103に格納する。長さsの値の決め方は、後述する。ある配列とオーバーラップする入力配列が存在するか否かを調べたいときには、この固定長部分配列テーブル103を参照する。参照の結果、ある入力配列の部分配列106が固定長ウィンドウ105の配列と完全に一致するとわかれば、同一クラスタに入るか否かをオーバーラップ部分の詳細な配列比較により検証する。そして、貪欲法に基づき、クラスタにメンバを逐次追加していく。

【００２４】

以下、本発明の方法について、詳細に説明する。処理の流れを、図２に示した。

まず、図２のステップ201に示すように、入力配列が配列長の降順に整列するようにソートする。これにより、例えば図３の配列304と配列305のオーバーラップ探索をするときに、配列304上に配列305の先頭又は末尾に一致する部分配列がないため、配列304と配列305を結合できないという状況が回避される。

【００２５】

次に、図２のステップ202に進み、固定長部分配列テーブル103を構築する。固定長部分配列テーブル103の構築に当たっては、図４に示すように、全ての入力配列101から両端の長さsの部分配列102を取り出しテーブル103に登録する。部分配列長sを長く取れば、入力配列間の真のオーバーラップに無関係に長さsの一致が偶然に発生する確率を低く抑えることができ、処理時間の短縮に繋がる。しかし、あまり部分配列長sを長くしすぎると、オーバーラップ探索の感度低下を招く。本発明では、処理時間短縮のため、sの値に下記の不等式（１）で表される下限を設けている。

【００２６】

【数１】

【数１】

$$s \geq \frac{1}{2}\log\frac{KN}{c} \qquad \cdots(1)$$

ただし、式（１）において、Nは入力配列数、Kは各配列から選択される部分配列の数、cはユーザにより与えられるパラメータであり、固定長部分配列テーブル103を一回参照する際に見つかる、入力配列間の真のオーバーラップと無関係な長さsの完全一致の数の期待値の上限を指定する量である。cを大きくすると、sを小さい値とすることができ、部分配列長が短くなるためオーバーラップ探索の感度を上げることができる。しかし、偶然の一致を処理する計算時間が増大するために処理速度は低下する。なお、本明細書では、対数の底は2とする。

【００２７】

固定長部分配列テーブル103に部分配列を登録するときに、その部分配列を含む入力配列を識別する情報と入力配列中の位置も同時に記録する。図４に示すような、これら部分配列、入力配列識別情報、入力配列中の位置の3つの値の組401それぞれを、エントリと呼ぶ。さらに、図４の402のような部分配列のそれぞれをキーと呼ぶ。固定長部分配列テーブル103の各エントリは、部分配列に一致する長さsの塩基配列をキーとして取り出すことができる。固定長部分配列テーブル103の実装には、AVL木などの、バランス木で2分木であるものを使用する（アルゴリズム辞典、共立出版、630頁）。

【００２８】

入力配列から部分配列を登録したあとで、固定長部分配列テーブル103中で、ユーザより与えられたパラメータF（後述）を上回る出現頻度を持つキーに対応するエントリを、全て削除する。この処理は、一般に核酸塩基配列にはしばしばリピート配列が含まれるため、入力配列間の真のオーバーラップに関係のない長さsの一致が多く見つかると予想されることから、出現頻度が極端に多いキーに対応するエントリを削除することを目的としている。

【００２９】

固定長部分配列テーブル103を構築した後、図２のステップ203に進み、個々のクラスタを構築していく。まず、最長の入力配列を選び、大きさ1のクラスタを構成する。ソートの工程201があるため、先頭の入力配列を選択するだけで、定数時間で容易に最長の入力配列を選ぶことができる。当該クラスタのコンセンサス配列104は、選択された最長の入力配列と同一の配列を複製して構築する。このコンセンサス配列上に幅sの固定長ウィンドウ105をとる。

【００３０】

図５を用いて、クラスタへ新規メンバを追加する方法について説明する。幅sの固定長ウィンドウ105を、その時点までに構築されているクラスタのコンセンサス配列104上を先頭から末尾まで移動させていく。移動中、固定長部分配列テーブル103を参照し、該クラスタのメンバとなる可能性のある入力配列の候補を探索する（図２のステップ204）。

【００３１】

ある位置で、固定長部分配列テーブル103を参照したところ、ある入力配列502と長さsの完全一致501が見つかったと仮定する。長さsの完全一致501が存在するだけでは、単なる偶然一致のことがあり、該クラスタにこの配列502を追加する条件として不充分である。オーバーラップ全体503が十分に類似しており、矛盾なくアセンブル可能であることを配列比較により確認する（図２のステップ205）。このときの配列比較は、コンセンサス配列及び入力配列が長さsの完全一致をもつ位置がわかっているため、Zhang, Z. et al., J. Comput. Biol., 7(1-2):203-14, 2000の高速アルゴリズムを利用する。

【００３２】

ステップ205の配列比較によってオーバーラップ全体503で配列が十分によく類似すると判断された場合、入力配列502を該クラスタに追加し、コンセンサス配列104も再構築して新しいコンセンサス配列504と置き換える（図２のステップ206）。コンセンサス配列が延長された部分505も、幅sの固定長ウィンドウ105の移動範囲に加える。クラスタに加えられた入力配列502に由来する、固定長部分配列テーブル103中のエントリは、削除する。

【００３３】

同様の処理を、コンセンサス配列104上に固定長ウィンドウ105の移動範囲が残っている間繰り返す。完成したクラスタは、順次、ファイル等に出力し、計算機の記憶装置に残さない。図６のように、未処理のまだどのクラスタにも属していない配列が残っている間、この処理を繰り返す。

以上が本発明の方法の主な流れである。上記の内容に加え、本発明の方法は、高速処理のために次のような特徴を持っている。

【００３４】

まず、コンセンサス配列104を格納するデータ構造には、双方向リストを用いる（図１４）。図１４に示す配列において、塩基T 1401と塩基G 1402の後に新たに塩基A 1501を挿入する必要が生じた場合、図１５に示すように、塩基T 1401と塩基G 1402の間に張られていたポインタを塩基A 1501へ張りなおし、塩基A 1501からは塩基T 1401と塩基G 1402へ新たなポインタを張ればよい。この処理は定数時間で可能である。もし、コンセンサス配列を図１６のような連続する記憶装置を使用するデータ構造、すなわちアレイを用いて実装すると、１塩基を挿入するために、図１７のように挿入したい位置以降の塩基を後ろにずらし、できた隙間1701に新しい塩基Aを挿入する必要があり、処理時間の平均は配列長に比例する量になってしまう。ただし、配列比較時には、ランダムアクセスを1回当たり定数時間で可能とするために、コンセンサス配列104のうち、比較する必要のある配列領域だけをアレイ状のデータ構造、すなわち連続した主記憶上の領域へコピーする。このときコピーされる配列領域の長さは、配列比較時に許容するギャップ数の最大値を２倍した値と、入力配列の配列長の和以下である。連続したメモリ領域では、配列中の塩基が格納されている位置が乗算と加算のみ計算可能であるため、任意の塩基に対して定数時間のランダムアクセスが実現される。

【００３５】

本発明の方法はまた、長さsの部分配列を少数の計算機ワードにエンコードすることで、定数時間で長さsの配列比較を行う。図７に、１計算機ワードが32bitである計算機の場合の例を示す。１計算機ワード701を使用すれば長さが16塩基まで、２計算機ワード702を使用すれば長さ32塩基までの配列をエンコードすることができる。A、T、G、C以外の文字が存在する場合は、強制的にA、T、G、Cの

いずれかの文字と同一と見なし、エンコードする。

【００３６】

本発明のクラスタリング及びアセンブル方法においては、さらに、核酸塩基配列に存在する可能性のあるシーケンシングエラーに対応するために、入力配列の両端だけでなく、入力配列上のより多くの部分配列を固定長部分配列テーブルに登録することで、僅かな塩基の誤りを許容したクラスタリングを可能とする。配列の端に近い部分配列が、オーバーラップ探索では重要となるため、配列両端からR塩基以内（Rはユーザパラメータ）の範囲から合計K個の部分配列を選択し、固定長部分配列テーブル103に登録する。この方法によると、ひとつの部分配列の塩基に誤りがある場合でも、他の部分配列によりオーバーラップを見つけることが可能になる。図１及び図４はK=2の例、図８はK=6の例である。

【００３７】

ここまで、本発明のクラスタリング及びアセンブル方法の概要について説明してきた。この方法が実際に高速なクラスタリング及びアセンブル方法であり、消費する計算機主記憶も少ないことを、理論的考察により示す。以下の説明では、次の記号を用いる。

N：入力配列の総配列数

D：入力配列の総塩基数

Li：あるクラスタiのクラスタリング完了時点でのコンセンサス配列長

Ni：あるクラスタiのクラスタリング完了時点でのクラスタメンバ数

L：クラスタリング完了時点での全クラスタのコンセンサス配列長の和

L’：クラスタリング完了時点での全クラスタ中最長のコンセンサス配列の配列長

n：クラスタリング完了時点でのクラスタ数

Dij：あるクラスタiにj番目に加えられた入力配列の配列長

M：最長の入力配列の配列長

E：一回の固定長部分配列テーブル103参照時に入力配列間のオーバーラップ部とは無関係であるにも関わらず偶然に見つかるエントリ数の期待値

K：ひとつの入力配列から固定長部分配列テーブル103に登録する部分配列の数

c：Eの上限を設定するためのユーザパラメータ

なお、計算時間や消費する主記憶の量を表現するために本明細書で用いる大文字のアルファベットOを用いた表記法については、平田富男著「アルゴリズムとデータ構造」森北出版株式会社（1990発行）"1.2節 計算量"に説明されている。

【００３８】

始めに、入力配列データが、A、T、G、C 4塩基のランダムな並びであることを仮定し、本発明の方法の高速性について説明する。まず、入力配列を配列長の降順にソートするために必要な時間はO(D+NlogN)である。なぜなら、配列長を求めるために、O(D)の時間が必要である。クイックソート又はマージソートを用いれば、ソートの処理はO(NlogN)で完了できる。

【００３９】

次に、本発明の方法が、固定長部分配列テーブル103を構成するために必要な時間はO(KNlogN)である。なぜなら、固定長部分配列テーブル103の実装に2分木のバランス木を使用しているので、全ての部分配列を登録するための計算時間は、O(KNlog(KN))であり、十分大きなNについてはN≧KよりO(KNlogN)である。また、頻度がFを上回るキーに対応するエントリの削除も、O(KNlogN)で行える。

さらに、本発明の方法が、i番目のクラスタひとつを構築するために必要な時間は式（２）で表される。

【００４０】

【数２】

【数２】

$$O\left((L_i+N_i)\log N+L_iE(M+\log N)+\sum_{j=1}^{N_i}D_{ij}+KN_i\log N\right)\quad\cdots(2)$$

なぜなら、i番目のクラスタひとつを構築する際の処理時間の詳細が、以下の通りとなるためである。

1．入力配列をひとつ選び、コンセンサス配列を構築する処理に要する計算時間は、その入力配列の長さのオーダーである。

【００４１】

２．固定長部分配列テーブル103を参照する処理の計算時間は、 O((Li+Ni－1+LiE)logN)である。なぜなら、まず、固定長部分配列テーブル103を一回参照する処理に要する時間がO(logN)である。さらに、参照回数の期待値について考察する。少なくともコンセンサス配列長に応じたO(Li)回の参照が必要である。入力配列間の真のオーバーラップに対応する一致がコンセンサス配列中の同一位置に集中した場合、さらにO(Ni－1)回の参照が必要になる。これらとは別に、偶然の一致が発見される回数の期待値がO(LiE)である。ゆえに、参照回数の期待値はO(Li+(Ni－1)+LiE)となる。

【００４２】

３．固定長部分配列テーブル103を参照し詳細な配列比較を行なった結果、クラスタに加えることが可能な入力配列が見つかった場合に、j番目のこうした入力配列について、前述の高速アルゴリズムによる配列比較及びコンセンサス配列更新にO(Dij)、固定長部分配列テーブル103の該入力配列に由来するエントリを削除する処理にO(KNi logN)の、合計O(Dij+KNi logN)の計算時間が必要である。

【００４３】

４．固定長部分配列テーブル103を参照した結果、オーバーラップ部に無関係な偶然の一致が見つかった場合、そうした偶然の一致が一回発見されるごとに、前述の高速アルゴリズムによる配列比較のためにO(M)の計算時間が必要である。

したがって、1つのクラスタを構築する際の工程全体の計算時間は、次式（３）の計算を経て、式（２）であるとわかる。

【００４４】

【数３】

【数３】

$$O(D_{i1})+O((L_i+(N_i-1)+L_iE)\log N)+\sum_{j=2}^{N_i}O(D_{ij}+K\log N)+O(L_iEM)$$

$$=O\left(D_{i1}+(L_i+N_i-1+L_iE)\log N+\sum_{j=2}^{N_i}D_{ij}+L_iEM+KN_i\log N\right)$$

$$=O\left((L_i+N_i)\log N+L_iE(M+\log N)+\sum_{j=1}^{N_i}D_{ij}+KN_i\log N\right)\quad\cdots(3)$$

ゆえに、全てのクラスタを計算する時間全体の計算時間は、次式（４）より、O((L+N)logN+LE(M+logN)+D+KN logN)である。

【００４５】

【数４】

【数４】

$$\sum_{i=1}^{n}\left[O\left((L_i+N_i)\log N+L_iE(M+\log N)+\sum_{j=1}^{N_i}D_{ij}+KN_i\log N\right)\right]$$

$$=O((L+N)\log N+LE(M+\log N)+D+KN\log N)\quad\cdots(4)$$

sが式（１）を満足するため NK/4^s≦cが成立すること、及び、配列がランダムであると仮定したことからE≦KN/4^s（4^sは、4のs乗を表す）であることの２点を考慮すれば、E≦cが成立する。

ゆえに、全てのクラスタを計算するために必要とされる計算時間は、O((L+N)logN+LE(M+logN)+D+KN logN)をL,N,KN≦Dを用いて変形すればO(D(M+logN))であるとわかる。

【００４６】

ソートに要する計算時間O(D+NlogN)、固定長部分配列テーブル103を構成するために要する計算時間O(DlogN)はいずれもO(D(M+logN))以下であるから、本発明のクラスタリング及びアセンブル方法全体の計算時間はO(D(M+logN))である。

仮に、全ての配列の長さが等しく、D=NMである場合、本発明の方法全体の計算時間はO(MN(M+logN))である。ここに、Mは配列長であり、Nに無関係な量である。したがって、本発明の方法の計算時間は、Nが増加するとき、Nの２乗よりも小さいオーダーでしか増加しない。すなわち、本発明の方法により、配列数の２乗よりも小さいオーダーでクラスタリング及びアセンブルを行うという課題が達成された。

【００４７】

一方、本発明の方法が消費する計算機主記憶の容量は、入力配列データ及び出力するクラスタ情報を除き、O(KN+L')である。なぜなら、本発明の方法が消費する計算機主記憶の容量は、入力配列データ及び出力するクラスタ情報を除くと、固定長部分配列テーブル103及び各時点で処理中のクラスタの情報だけである。固定長部分配列テーブル103を格納するために必要とされる主記憶の容量は、２分木を用いた場合O(KN)であり、処理中クラスタ情報は、各メンバの配列長の和で抑えられるからO(L')である。これ以外に必要な計算機主記憶は、O(1)である。O(KN+L')は、入力配列の長さに依存せず、配列数に比例するオーダーでしか増加しない量である。

【００４８】

図１８は、上記の方法を実行する核酸塩基配列アセンブル装置の構成例を示す図である。図１８に示したように、本発明の核酸塩基配列アセンブル装置は、計算を行うCPU 1801、インタフェースを実現するためのディスプレイ1802、キーボード1803、ポインティングデバイス1804を備え、入力配列を配列長の降順にソートするプログラム1805、固定長部分配列テーブル103を構築するプログラム1806、コンセンサス配列と一致する部分配列をもつ入力配列を探索するプログラム1807、コンセンサス配列と入力配列がアセンブル可能かを判断するプログラム1808、コンセンサス配列の再構築を行うプログラム1809を格納し、さらに固定長部分配列テーブル103を格納する主記憶装置1810、入力配列1811が格納され、さらに

クラスタリング及びアセンブル結果1812を格納することができる補助記憶装置1813から構成される。

【００４９】

ディスプレイ1802、キーボード1803、ポインティングデバイス1804を使用し、入力配列及び本発明の方法が必要とするパラメータを指定した後、CPU 1801が主記憶装置1810に格納されたプログラムを実行し、本発明の方法によりクラスタリング及びアセンブルが行なわれる。入力配列1811は補助記憶装置1813より読みこまれる。出力されるクラスタリング及びアセンブル結果1812は、補助記憶装置1813に格納することができる。本発明の方法によりクラスタリング及びアセンブルが行なわれている間、ディスプレイ1802に処理経過が表示され、また、処理終了後には、ディスプレイ1802に処理結果を表示させることができる。

【００５０】

以下、ディスプレイ1802、キーボード1803、ポインティングデバイス1804により実現される、パラメータ設定インタフェース、経過表示インタフェース、結果表示インタフェースと、それらを呼び出すためのメインインタフェースを説明する。

パラメータ設定インタフェースを用いて、各入力配列から固定長部分配列テーブル103に記録する固定長部分配列の数K、固定長部分配列の位置の配列両端からの距離の上限R、それぞれの固定長部分配列の位置、固定長部分配列テーブル103参照時に真のオーバーラップと無関係に見つかる偶然キーが一致するエントリの数の期待値の上限c、固定長部分配列長s、固定長部分配列テーブル103に頻繁に出現する部分配列の頻度の上限Fを入力する。また、経過表示インタフェースを用いてクラスタリング及びアセンブル処理中の処理済み配列の数及び入力配列全体に対する比率、構成済みクラスタ数、クラスタ要素数の平均、コンセンサス配列への各クラスタメンバ配列のアセンブル状況、アセンブル時に固定長部分配列テーブル103を使用することで得られた完全マッチの配列、アセンブル時のオーバーラップ長を表示する。処理終了後、ユーザが指定した入力配列又はクラスタに対して、処理中と同様の情報を結果表示インタフェースにより表示する。

【００５１】

図９、図１０、図１１、図１２を用いて、本発明におけるユーザインタフェースの一例を詳細に説明する。

図９を用いて、メインインタフェースの一例について説明する。このメインインタフェース901は、入力配列選択部907、パラメータ入力インタフェースを出現させるためのパラメータ設定ボタン905、及び経過表示インタフェースを出現させ、クラスタリング及びアセンブルを実行するためのアセンブル実行ボタン906を有する。

【００５２】

ユーザはまず、入力配列を格納したファイルのパスをファイルパス入力エリア902に入力する。この例では、参照ボタン903をポインティングデバイス1804でクリックすることによりファイルダイアログを出現させることができる。そのファイルダイアログを用いて、入力配列を格納したファイルを選択してもよい。入力配列のファイルパスが入力されると、入力データの配列数Nが計算され、配列数表示エリア904に表示される。

入力配列を指定した後、パラメータ設定ボタン905をクリックすることにより、パラメータ設定インタフェースが現れる。パラメータ設定インタフェースの例については、後述する。

【００５３】

アセンブル実行ボタン906をクリックすると、経過表示インタフェースが現れ、クラスタリング及びアセンブルの処理が開始される。アセンブル実行ボタン906は、入力配列が入力されるまではクリックすることができない。経過表示インタフェースの例は、後述する。処理の終了後、経過表示インタフェースは自動的に閉じられ、結果表示インタフェースが現れる。結果表示インタフェースの例は、後述する。

【００５４】

図１０を用いて、パラメータ設定インタフェースの例について説明する。図示したパラメータ設定インタフェース1001は、固定長部分配列位置選択部1021、固定長部分配列長設定部1022、固定長部分配列キー頻度上限入力部1023を有する。

まず、ユーザはクラスタリング及びアセンブル時に各配列から抽出する固定長

部分配列数Kを、固定長部分配列位置選択部1021の入力エリア1002にキーボード1803等を用いて入力し指定することができる。さらに、固定長部分配列と配列の5'端又は3'端からの距離の上限を決定するパラメータRを入力エリア1003に入力し指定することができる。Rの値は、グラフィカルユーザインタフェース1004内のスライダー1005を横に動かすことによっても指定可能である。ボックス1006は、固定長部分配列を抽出する位置を表し、1008又は1009で指定される固定長部分配列長に比例する幅をもつ。グラフィカルユーザインタフェース1004内には、入力エリア1002で指定されたKの値に等しい数のボックス1006が表示される。ボックス1006は、入力配列を表す線分1007上を両端からR塩基の範囲内で、ポインティングデバイス1804を使用し、ボックス同士が重なり合わない限り自由に動かすことができる。配列先頭付近のボックス1006を配列先頭からR塩基以上離そうとすると、そのボックス1006は配列末尾からR以内の位置に移動する。逆の場合は同様に、配列先頭からR塩基以内の位置にボックス1006が移動する。

【００５５】

また、ユーザは固定長部分配列テーブル103参照時に真のオーバーラップと無関係に見つかる偶然にキーが一致するエントリ数の期待値上限cの値を、固定長部分配列長設定部1022の入力及び表示エリア1008に入力し指定することが可能である。cが入力されると、固定長部分配列長sが式（１）を満足する最小の整数の値として自動的に計算され、入力及び表示エリア1009に表示される。固定長部分配列長sを直接、固定長部分配列長の入力及び表示エリア1009に入力し指定することも可能である。sが入力されると、式（１）に矛盾しない最小のcの値、すなわち$NK/4^{s} \leqq c$を満足する最小のcの値が自動的に計算され、入力及び表示エリア1008に表示される。

さらに、固定長部分配列テーブル103におけるキー頻度の上限Fを、固定長部分配列キー頻度上限入力部1023を用いて指定する。

【００５６】

まず、Fの値は、入力及び表示エリア1011に数値として直接入力し指定することが可能である。キー頻度がFを超えるエントリを削除する処理を行わない場合には、チェックボックス1012にチェックをつける。これ以外に、Fの値を指定す

る手段として、固定長部分配列テーブル103を作成した後で、実際の固定長部分配列出現頻度を見ながらFの値を設定することもできる。グラフ表示エリア1013には、ファイルパス入力エリア902で指定された入力配列に基づき、固定長部分配列テーブル103を構築した場合に、横軸を出現頻度、縦軸を出現頻度の順位とするグラフが表示される。縦方向の拡大率は、スライダー1017で変更できる。このグラフ上で、Fを表す線分1014を動かすことでFの値を設定可能である。一方、表示エリア1015には、キーである固定長部分配列と固定長部分配列テーブル103中の出現頻度の組を出現頻度の降順に整列した表が表示される。表示エリア1015上でFの値を表す線分1016を動かしても、やはりFの値を設定することができる。Fの上限値入力及び表示エリア1011、Fを表す線分1014、1016の3つのうち、1つが操作されFの値が変更されると、残りの2つの表示も新しいFの値に応じて更新される。

設定したパラメータの値を確定させ、クラスタリング及びアセンブルに使用する場合は、ボタン1018をクリックする。設定したパラメータの値を破棄し、パラメータ入力インタフェース表示前の状態に戻すには、ボタン1019をクリックする。

【００５７】

次に、図１１を用いて、クラスタリング及びアセンブルの処理経過を表示する経過表示インタフェースの一例について説明する。この例の経過表示インタフェース1101は、全体処理状況表示部1121、クラスタアセンブル状況表示部1122、配列比較状況表示部1123を有する。

全体処理状況表示部1121には、処理中、各時点までにいずれかのクラスタに加えられた配列の数が表示エリア1102に、生成されたクラスタ数が表示エリア1103に、クラスタ要素数の平均が表示エリア1104に表示される。棒グラフ1105内に、処理済み配列数に相当する部分1106が色を変えるなど容易に視認できる手段を用いて表示される。

【００５８】

クラスタアセンブル状況表示部1122には、個々のクラスタのアセンブル状況が表示される。コンセンサス配列104は横長の長方形1107として表示する。長方形1

107中、固定長ウィンドウ105が既に走査した領域に相当する領域1108を、色を変えるなど容易に視認できる手段で表示する。固定長ウィンドウ105がコンセンサス配列104上を走査した際、各位置でその位置の長さsの配列をキーとして固定長部分配列テーブル103を参照したときに、見つかった長さsの完全マッチの数をグラフ1109として表示する。極端にマッチ数が多い領域は、リピート配列又は機能ドメインの存在を示唆している可能性がある。

【００５９】

コンセンサス配列104にアセンブルされた入力配列502は、水平の線分1110として表示される。これらの入力配列をクラスタに加える際に使用された長さsの完全一致に相当する領域は、色を変えるなどして容易に視認できるようにした領域1111として表示される。ただし、クラスタに最初に加えられた配列は、長さsの完全一致に基づきクラスタへ加えられたわけではないため、その表示1112は長さsの完全一致の表示1111を伴なわない。

これら以外に、クラスタに新規メンバを追加する際に、長さsの完全一致に相当する領域の塩基配列を配列比較状況表示部1123の表示エリア1113に表示し、アセンブル時のオーバーラップ長を表示エリア1114に表示する。

【００６０】

表示のオーバーヘッドに伴う速度低下を防ぐため、「同時表示」と書かれたトグルボタン1115をクリックするごとに、表示を停止するか否かを切り換えることができる。「一時停止」と書かれたボタン1116を押している間、表示、クラスタリング及びアセンブルの処理を一時的に中断させることができる。

【００６１】

図１２を用いて、クラスタリング及びアセンブル処理結果を表示する結果表示インタフェースの一例について説明する。図示した結果表示インタフェース1201は、全体の処理結果を表示する表示エリア1222、クラスタアセンブル状況を表示する表示エリア1223、配列アセンブル状況を表示する表示エリア1224、表示クラスタを選択するエリア1225を有する。

全体の処理結果を表示する表示エリア1222には、総入力配列数、生成されたクラスタ数、平均クラスタサイズが表示される。

【００６２】

生成された各クラスタのアセンブル状況を表示する表示エリア1223について説明する。コンセンサス配列104の表示1204、固定長部分配列テーブル参照時に見つかった長さsの完全一致の頻度を表すグラフ1205、クラスタメンバである入力配列502を表す線分1206、該入力配列502をクラスタに加える際に使用された長さsの完全一致1207は、それぞれ経過表示インタフェース1101における表示1107、1109、1110、1111と同様である。ただし、経過表示インタフェース1101では配列アセンブル状況として処理中の入力配列についてのみ、完全一致の配列及びオーバーラップ長を表示エリア1113、1114に表示していたが、結果表示インタフェース1201では、表示中のクラスタの任意の配列を選択し、オーバーラップ長1209と長さsの完全一致の配列1210を表示させることができる。その配列は、枠1208のような手段により強調表示される。配列を表す線分1206をポインティングデバイス1804でクリックするか、キーボード1803で該配列を他の配列に変更することができる。

【００６３】

結果表示インタフェース1201では、表示するクラスタも選択することができる。入力配列名のエリア1211に入力配列名を入力した上で、表示ボタン1212をクリックすることにより、その入力配列を含むクラスタのアセンブル状況が表示される。表示の際、ユーザによって指定された入力配列は枠1208のような手段で強調表示され、その入力配列について、オーバーラップ長と長さsの完全一致の配列がそれぞれ表示エリア1209、1210に表示される。入力配列だけでなく、クラスタを指定して表示させることもできる。クラスタリング及びアセンブルの処理中、出力するクラスタに通し番号をつけ、さらに、その番号を入力エリア1213に入力し表示ボタン1214をクリックすることで該クラスタの表示を行う。

【００６４】

実際に、本発明の方法に基づき、入力配列をクラスタリング及びアセンブル処理を行うソフトウェアを実装した。ただし、この実装ではコンセンサス配列を表現するデータ構造にアレイを使用したため、本発明の方法よりも漸近的時間計算量が大きくなっている。また、固定長部分配列テーブル103の実装には、C++言語

のライブラリSTLのmultimapクラスを使用した。このデータ構造はバランス木と同様に、要素の挿入・検索・削除を要素数の対数に比例する時間で処理可能である。この実装は、グラフィカルなインタフェースを含んでいない。

【００６５】

クラスタリング及びアセンブル処理を行うためには、sの値を設定する必要がある。真のオーバーラップでない、偶然の一致で見つかるエントリ数の期待値は、配列がランダムであると仮定すればNK/4^s以下であり、sが大きいほど計算時間は短くなるが、配列にエラーが存在し完全一致が成立しなくなる可能性を小さくするために、sは可能な限り小さいことが望ましい。1計算機ワードが32bitの計算機で固定長部分配列を1計算機ワードで表現するためには16塩基、2計算機ワードで表現するためには32塩基以内の固定長部分配列を使用する必要がある。

【００６６】

最適なsの値を調べるため、N，K，sを変化させ、クラスタリング及びアセンブルに要する時間を計測した。対象とした配列データは、mRNAから得られたESTのクラスタリング及びアセンブルをシミュレートするために、一般的なmRNA長に近い長さ2000塩基のランダムな配列を、生体内のタンパク質数といわれる10万配列用意し、そこからランダムにEST長と同程度の長さ500塩基の配列を抽出して作成した。計算機は、CPUのクロック周波数が1.7GHz、主記憶容量が1GBのものを使用した。結果を、表１に示す。表１は、クラスタリング及びアセンブルに要した時間、sを減少させた場合の処理時間の増加率（固定長部分配列の長さがs+1の場合の処理時間を、長さがsでの処理時間で除した値）、固定長部分配列テーブル参照時の真のオーバーラップに関係なく偶然見つかるエントリ数の期待値NK/4^sからなる。

【００６７】

【表１】

表1

クラスタリング及びアセンブル処理に要した時間(秒)													
						固定長部分配列の長さ(s)							
N	K	20	19	18	17	16	15	14	13	12	11	10	9
65536	2	142	127	125	123	122	122	122	123	126	142	201	454
65536	8	156	154	150	151	149	148	150	153	166	222	450	1542
262144	2	526	478	474	469	466	465	466	481	535	753	1733	9297
262144	8	581	579	570	566	570	563	578	622	810	1575	5484	-
1048576	2	1537	1542	1526	1511	1513	1517	1553	1694	2293	5633	-	-
固定長部分文字列長が減少した際の処理時間の増加率													
N	K	20	19	18	17	16	15	14	13	12	11	10	9
65536	2	-	0.8943	0.9842	0.984	0.9918	1	1	1.0081	1.0243	1.1269	1.4154	2.2587
65536	8	-	0.9871	0.974	1.0066	0.9867	0.9932	1.0135	1.02	1.0849	1.3373	2.027	3.4266
262144	2	-	0.9087	0.9916	0.9894	0.9936	0.9978	1.0021	1.0321	1.1122	1.4074	2.3014	5.3646
262144	8	-	0.9965	0.9844	0.9929	1.007	0.9877	1.0266	1.0761	1.3022	1.9444	3.4819	-
1048576	2	-	1.0032	0.9896	0.9901	1.0013	1.0026	1.0237	1.0907	1.3536	2.4566	-	-
固定長部分文字列の偶然一致の期待値(NK/4^s)													
N	K	20	19	18	17	16	15	14	13	12	11	10	9
65536	2	0	0	0	0.00001	0.00003	0.00012	0.00049	0.00195	0.00781	0.03125	0.125	0.5
65536	8	0	0	0.00001	0.00003	0.00012	0.00049	0.00195	0.00781	0.03125	0.125	0.5	2
262144	2	0	0	0.00001	0.00003	0.00012	0.00049	0.00195	0.00781	0.03125	0.125	0.5	2
262144	8	0	0.00001	0.00003	0.00012	0.00049	0.00195	0.00781	0.03125	0.125	0.5	2	8
1048576	2	0	0.00001	0.00003	0.00012	0.00049	0.00195	0.00781	0.03125	0.125	0.5	2	8

【００６８】

この表より、sが大きいときには計算時間はほとんど変化しないが、ある程度以上sが小さくなると、急激に計算時間が増加することがわかる。例えば、N=65536，K=2のときには、s=9のときはs=10に比べ2倍以上の計算時間を要している。計算時間を抑制するためには、sの値を処理時間の急激な増加が見られない範囲内とすることが望ましい。表１より、この実験で扱った配列数のデータを扱う際に、計算時間の増加を2倍以内とするためには、sを少なくとも10以上とする必要があることがわかる。更に詳細に表１について考察すると、$NK/4^s \leqq 0.125$のとき、計算時間の増加率が1.5以下であることがわかる。そこで、c=0.125とし、sの値を式（１）を満足するように取ることが、計算時間の抑制に有効であることがわかる。

【００６９】

上記ソフトウェアを用いて、mRNA由来の核酸塩基配列のクラスタリングを試みた。使用したデータは、米国公共機関のデータベースで公開されている配列データで、総配列数は1,536,220、総塩基数は656,663,661であった。計算機は、CPUのクロック周波数が450ＭＨｚ、主記憶容量が4GBの計算機を使用した。固定長部

分配列長sはc=0.125及び式（１）より13とした。また、K=2，R=13（配列の先頭及び末尾からs=13の長さの部分配列を固定長部分配列テーブル103に登録）、F=10とした。

クラスタリング及びアセンブル処理は、172分57秒で完了した。得られたクラスタの数は732,166であった。

【００７０】

前述のHuang，X. and Madan，A.，Genome Research,9:868-877，1999の方法を実装したソフトウェアが開発されているが、入力配列数の制限から、100万配列を同時に処理することはできない。一方、Altschul，S.F. et al.，Nucleic Acid Research，25:3389-3402,1997の方法は、クラスタリング及びアセンブルの処理全体を含む方法ではないが、クラスタリング及びアセンブルの処理の一部である配列間の全組み合わせのオーバーラップ探索を行なうことが可能である。しかし、ひとつの配列とオーバーラップする配列をすべて探索するために、ＣＰＵのクロック周波数が450MHzのワークステーションで9秒程度の時間を要し、1,536,220配列のオーバーラップをすべて探索すると160日程度かかることが予想される。

本発明の方法よりも漸近的時間計算量が大きな前記ソフトウェアで、172分57秒という短時間で1,536,220配列からなる配列データのクラスタリング及びアセンブルに成功したことにより、本発明の方法の有効性が実証された。

【００７１】

【発明の効果】

本発明によれば、クラスタリング及びアセンブル処理を、Dを入力配列の総塩基数、Nを総入力配列数、Mを最長の入力配列の長さとするときO(D(M+logN))の以下の計算時間で行うことが可能となり、150万配列を超える莫大な数の配列データのクラスタリングが数時間で可能となる上、グラフィカルなユーザインタフェースが提供される。

【図面の簡単な説明】

【図１】

本発明の基本的なアイデアを表す説明図。

【図２】

本発明のクラスタリング及びアセンブル方法の、全体的な流れを表すフローチャート。

【図３】

先頭及び末尾の部分配列を固定長部分配列テーブルに登録する理由を説明する図。

【図４】

固定長部分配列テーブル構築方法の説明図。

【図５】

クラスタに新規メンバを追加するときの、配列オーバーラップ状況の説明図。

【図６】

本発明の方法の進行過程を表す説明図。

【図７】

固定長部分配列の計算機ワードへのエンコーディングの例を表す説明図。

【図８】

固定長部分配列テーブル作成時に、各配列の先頭及び末尾から複数の部分配列抽出する方法の説明図。

【図９】

メインインタフェースの一例の説明図。

【図１０】

入力インタフェースの一例の説明図。

【図１１】

経過表示インタフェースの一例の説明図。

【図１２】

結果表示インタフェースの一例の説明図。

【図１３】

mRNA由来のESTのクラスタリング及びアセンブル処理の概要の説明図。

【図１４】

データ構造に双方向リストを用いて塩基配列を格納する方法の説明図。

【図１５】

データ構造に双方向リストを用いて格納された塩基配列に新たな塩基を挿入する方法の説明図。

【図１６】

データ構造にアレイを用いて塩基配列を格納する方法の説明図。

【図１７】

データ構造にアレイを用いて格納された塩基配列に新たな塩基を挿入する方法の説明図。

【図１８】

本発明による核酸塩基配列アセンブル装置の構成例を示す図。

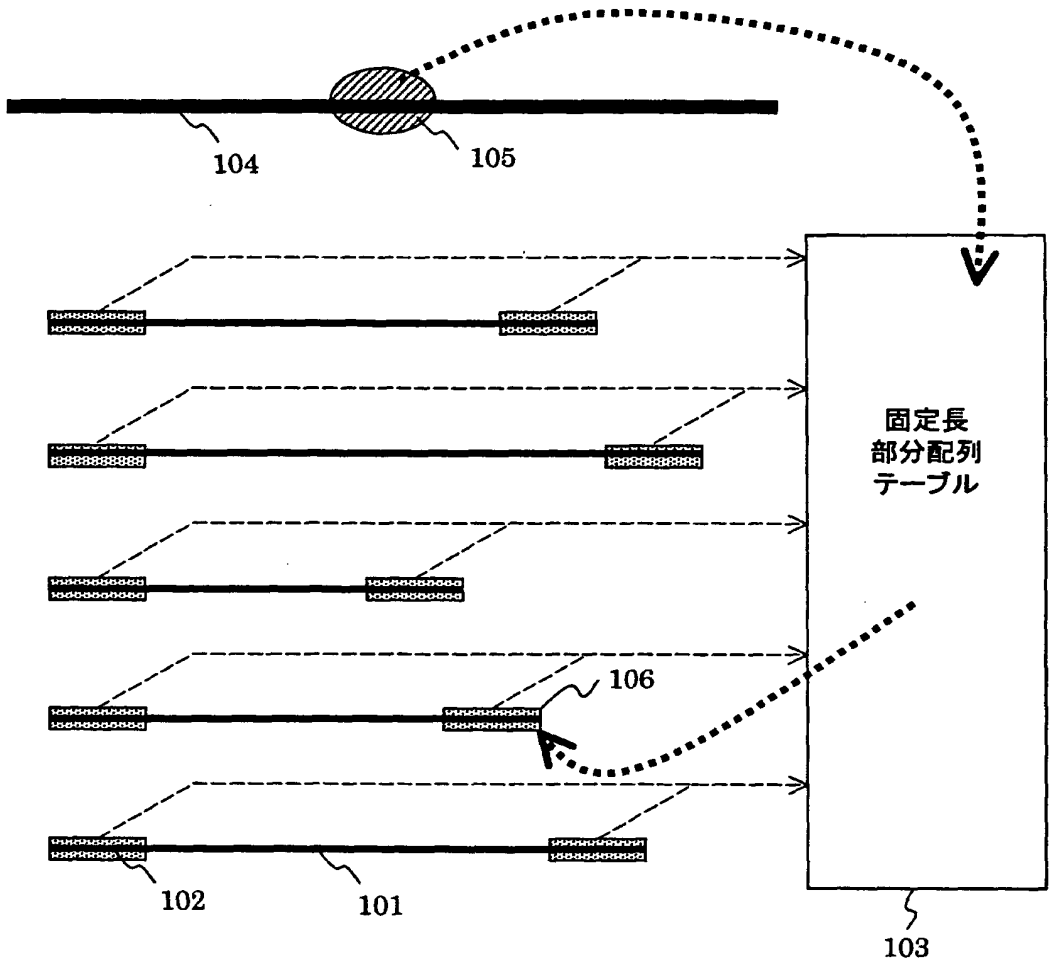
【符号の説明】

１０１：クラスタリング及びアセンブルの対象となる入力配列、１０２：入力配列から選択された長さsの固定長部分配列、１０３：固定長部分配列テーブル、１０４：ある時点で処理対象としているクラスタのコンセンサス配列、１０５：固定長ウィンドウ、３０１～３０５：入力配列、３０３，３０６：２つの入力配列のオーバーラップ部分、４０１：エントリ、４０２：キー、５０２：入力配列、５０３：コンセンサス配列と入力配列のオーバーラップ部、５０４：新しいコンセンサス配列、５０５：古いコンセンサス配列104に無く、新しいコンセンサス配列504に存在する配列領域、１３０１：mRNA、１３０２：EST、１３０３：EST配列集合、１３０５：配列間がオーバーラップしているペア、１３０６：ESTのアセンブルにより得られた塩基配列

【書類名】　　図面

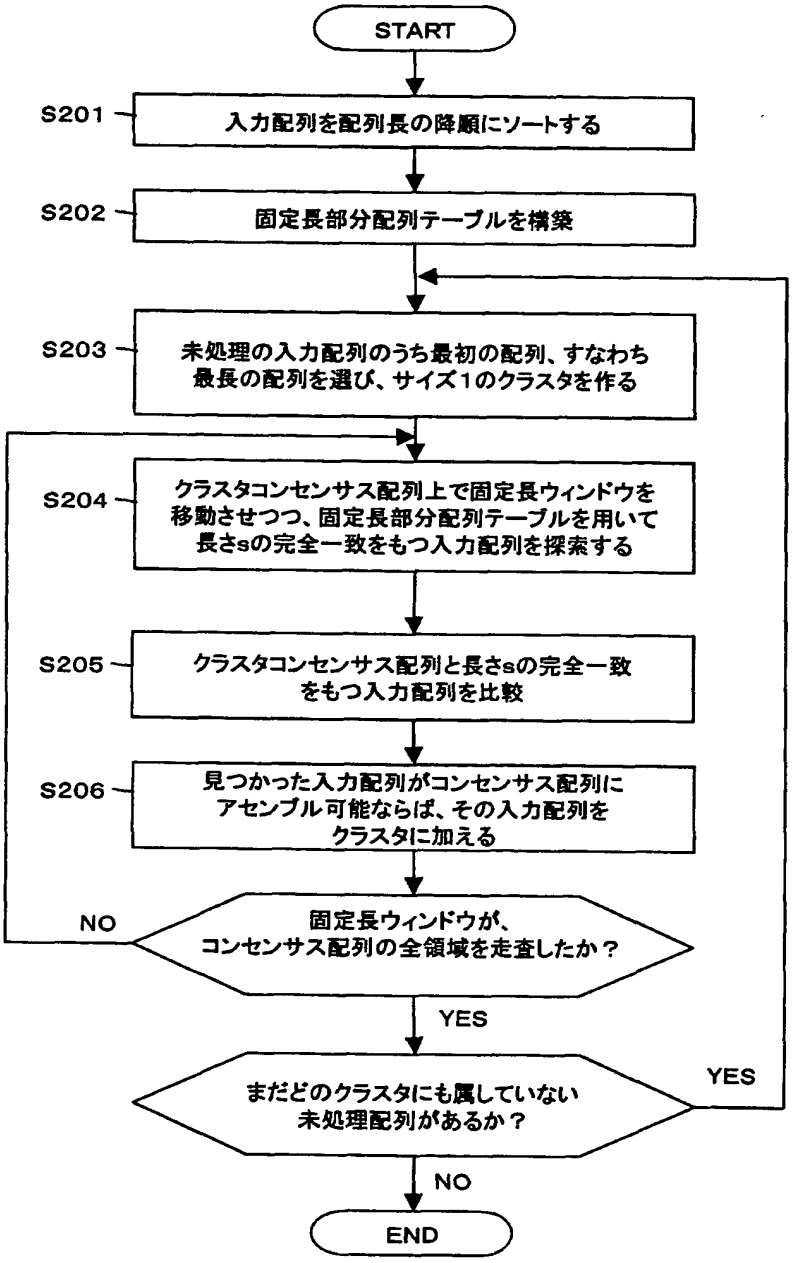
【図１】

図１



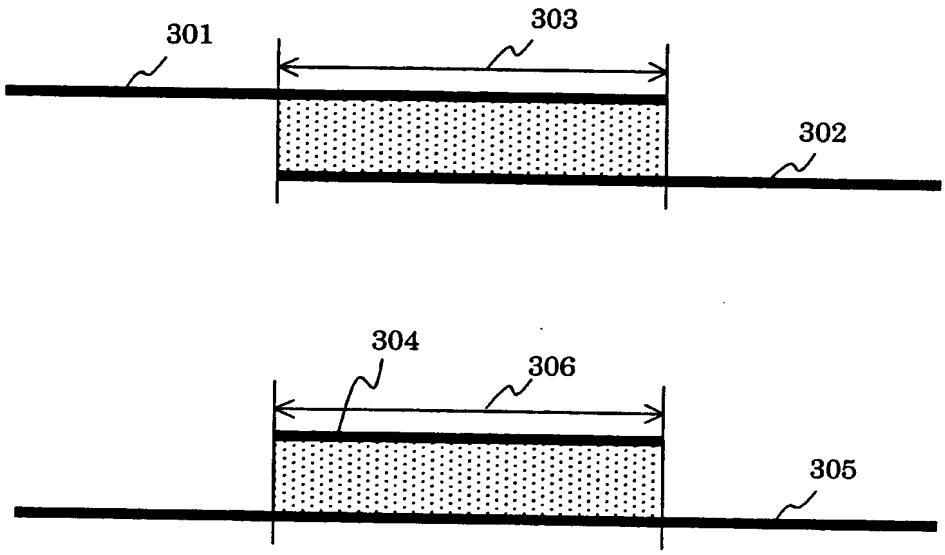
【図２】

図2


START
S201
入力配列を配列長の降順にソートする
S202
固定長部分配列テーブルを構築
S203
未処理の入力配列のうち最初の配列、すなわち
最長の配列を選び、サイズ1のクラスタを作る
S204
クラスタコンセンサス配列上で固定長ウィンドウを
移動させつつ、固定長部分配列テーブルを用いて
長さsの完全一致をもつ入力配列を探索する
S205
クラスタコンセンサス配列と長さsの完全一致
をもつ入力配列を比較
S206
見つかった入力配列がコンセンサス配列に
アセンブル可能ならば、その入力配列を
クラスタに加える
NO
固定長ウィンドウが、
コンセンサス配列の全領域を走査したか？
YES
YES
まだどのクラスタにも属していない
未処理配列があるか？
NO
END

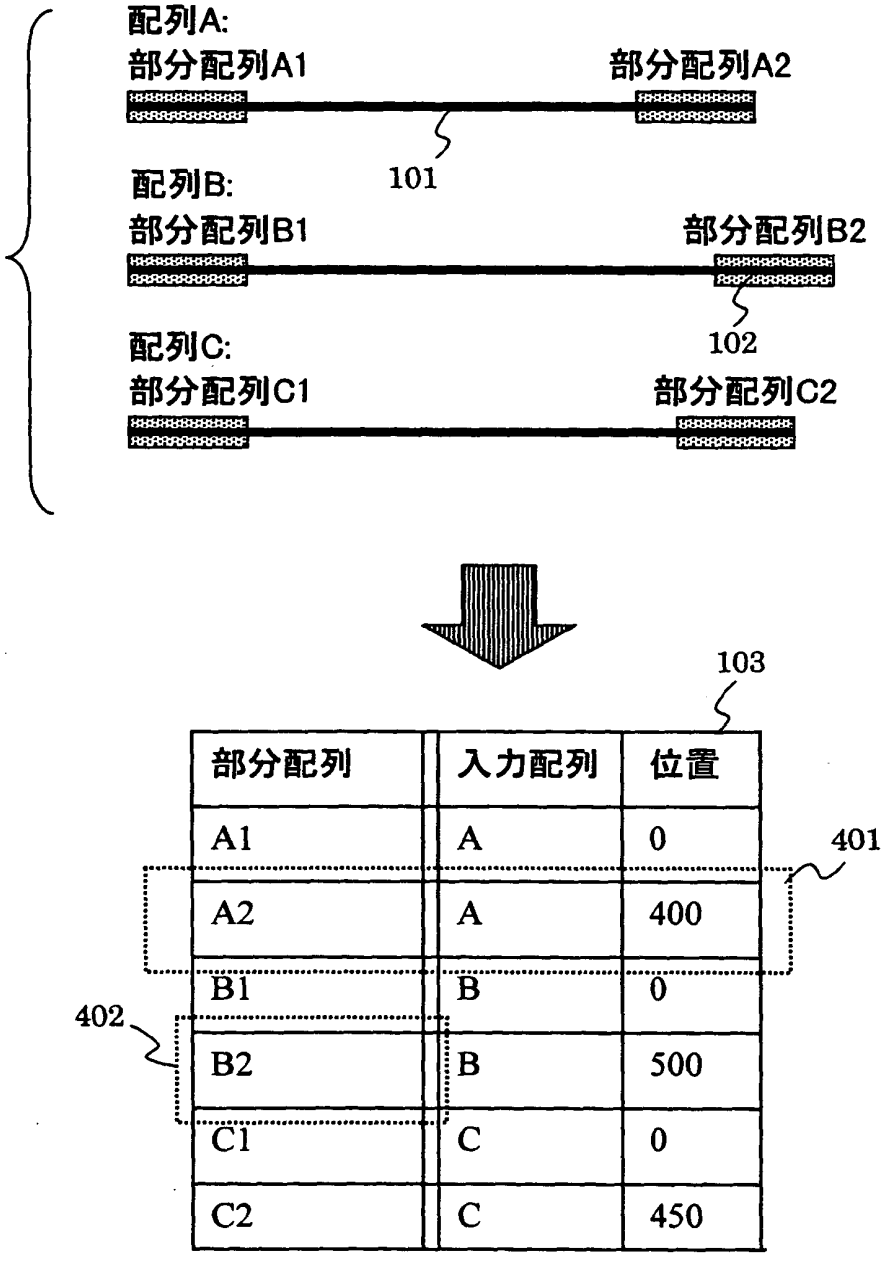
【図３】

図3



【図４】

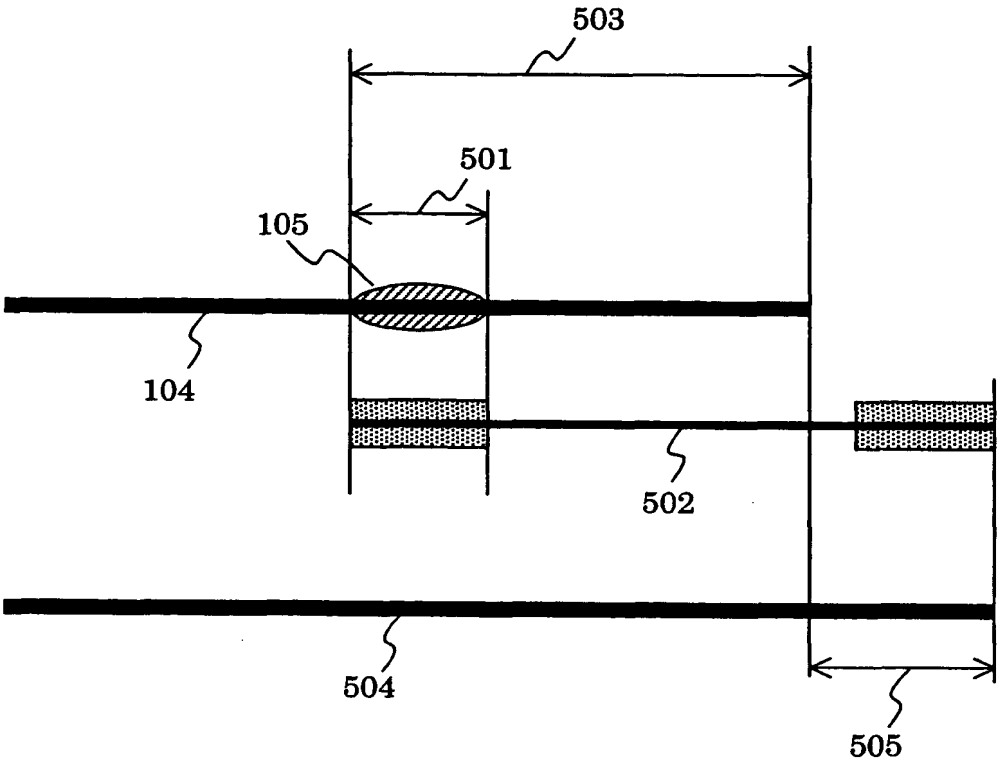
図4



部分配列	入力配列	位置
A1	A	0
A2	A	400
B1	B	0
B2	B	500
C1	C	0
C2	C	450

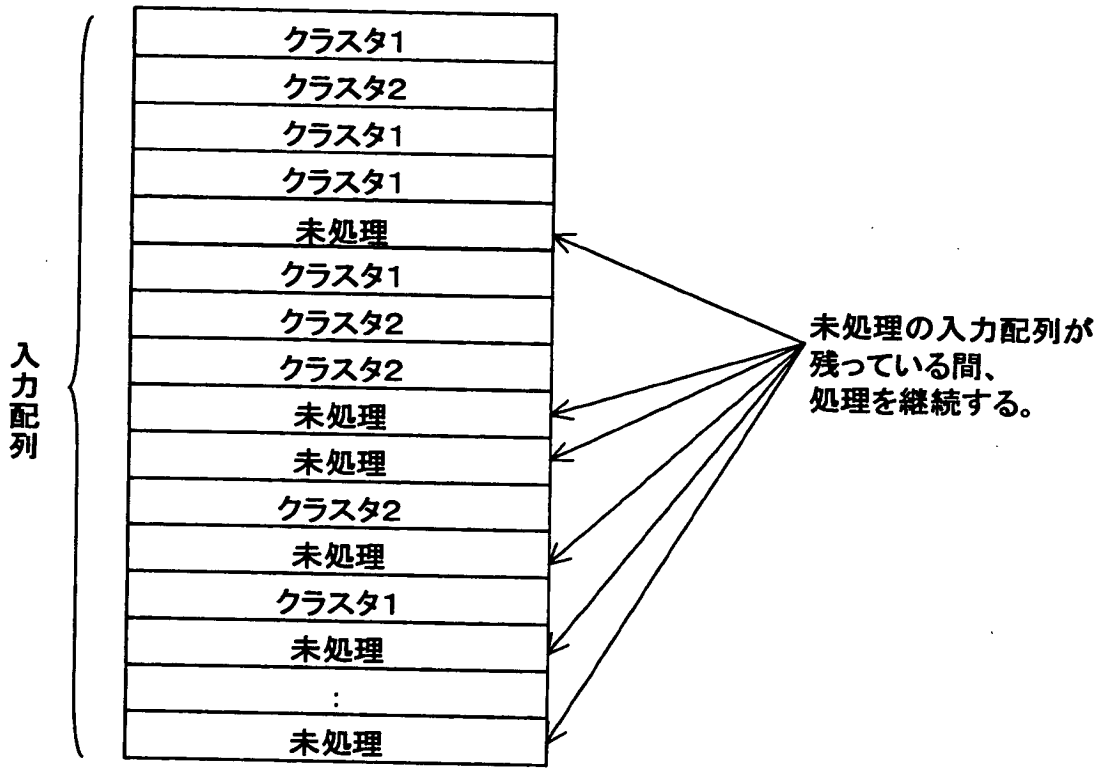
【図５】

図５



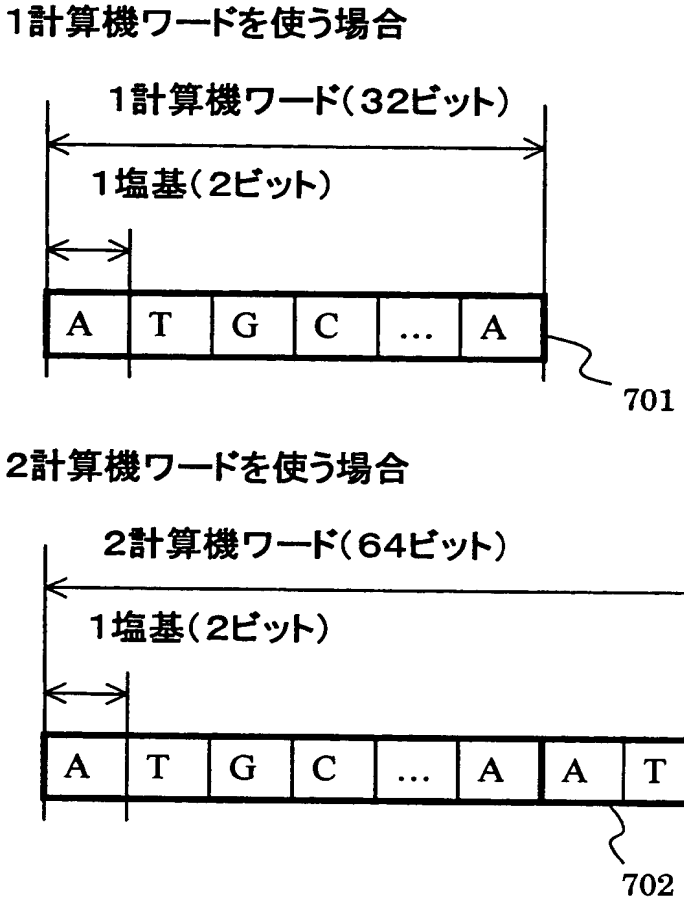
【図６】

図6


入力配列
クラスタ1
クラスタ2
クラスタ1
クラスタ1
未処理
クラスタ1
クラスタ2
クラスタ2
未処理
未処理
クラスタ2
未処理
クラスタ1
未処理
:
未処理
未処理の入力配列が
残っている間、
処理を継続する。

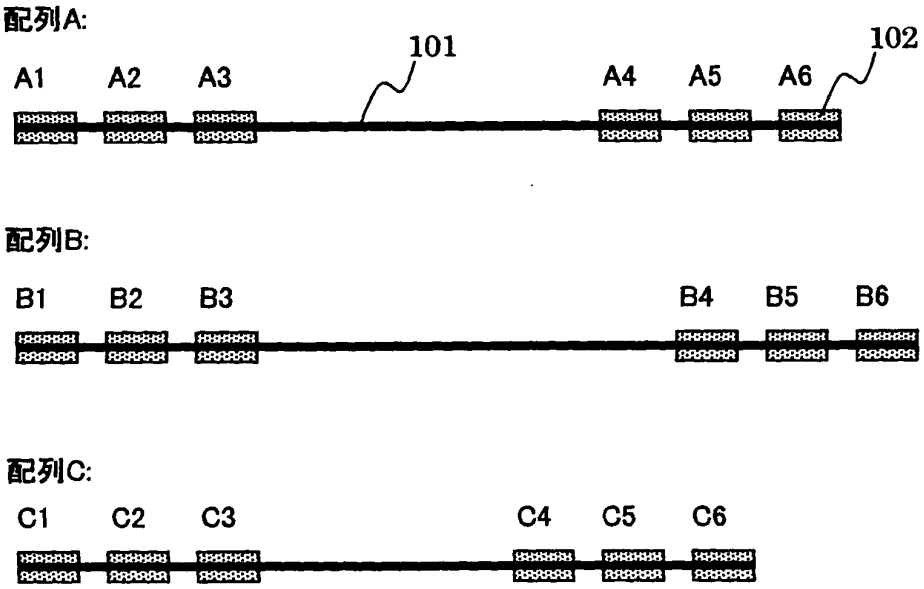
【図７】

図7


1計算機ワードを使う場合
1計算機ワード(32ビット)
1塩基(2ビット)
A
T
G
C
…
A
701
2計算機ワードを使う場合
2計算機ワード(64ビット)
1塩基(2ビット)
A
T
G
C
…
A
A
T
G
C
…
A
702

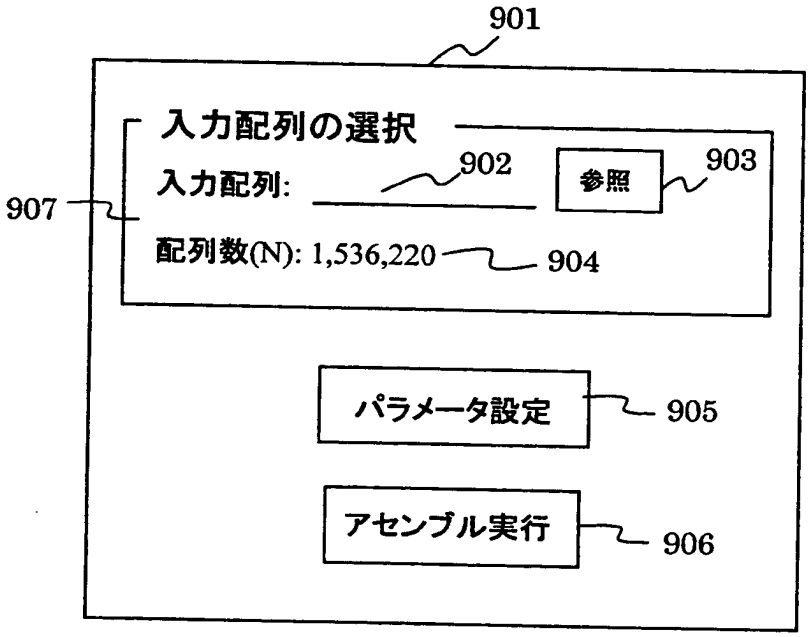
【図８】

図8


配列A:
A1
A2
A3
101
A4
A5
A6
102
配列B:
B1
B2
B3
B4
B5
B6
配列C:
C1
C2
C3
C4
C5
C6

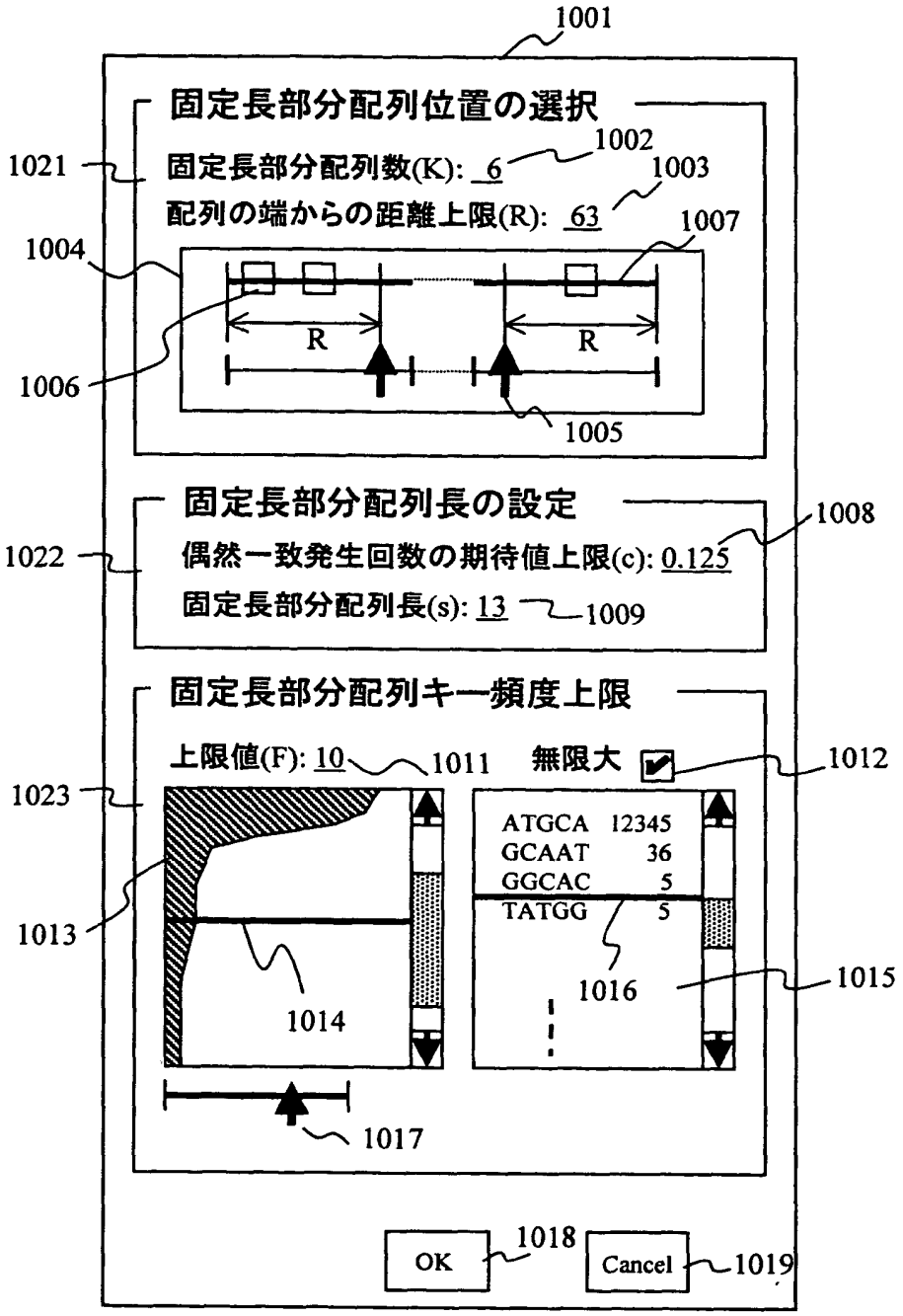
【図９】

図9


901
入力配列の選択
入力配列:
902
参照
903
907
配列数(N): 1,536,220
904
パラメータ設定
905
アセンブル実行
906

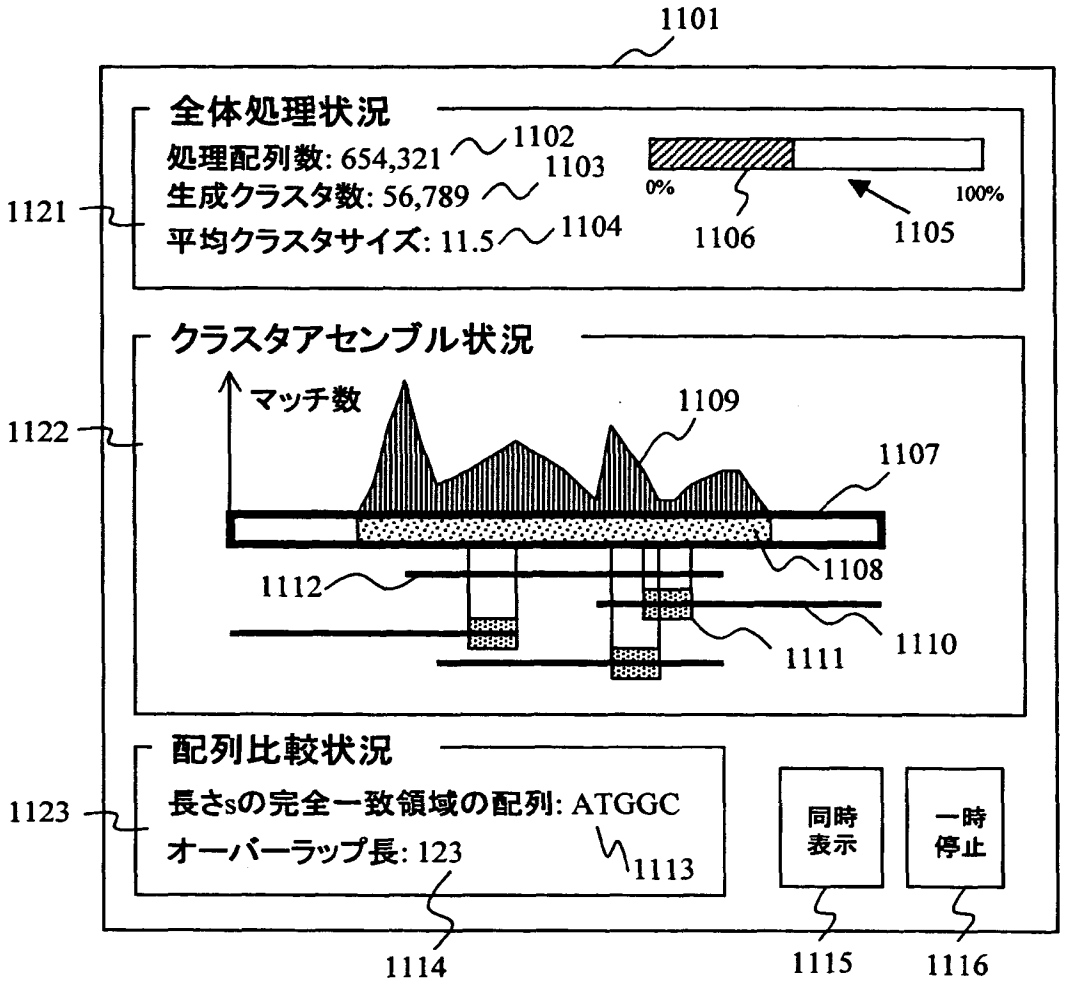
【図１０】

図10


1001
固定長部分配列位置の選択
1021
固定長部分配列数(K): 6
1002
1003
配列の端からの距離上限(R): 63
1007
1004
R
R
1006
1005
固定長部分配列長の設定
1008
1022
偶然一致発生回数の期待値上限(c): 0.125
固定長部分配列長(s): 13
1009
固定長部分配列キー頻度上限
上限値(F): 10
1011
無限大
1012
1023
ATGCA 12345
GCAAT 36
GGCAC 5
TATGG 5
1013
1014
1016
1015
1017
OK
1018
Cancel
1019

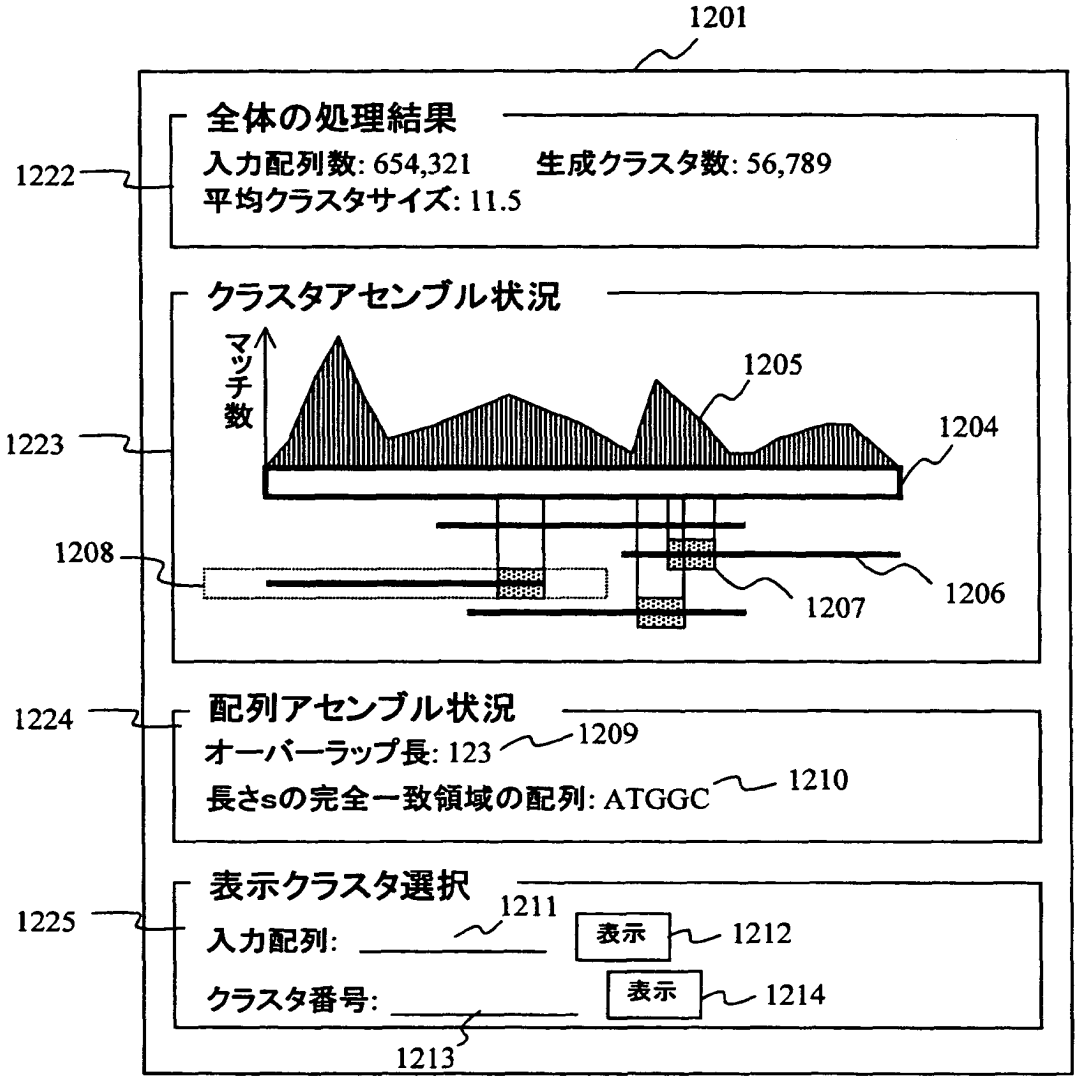
【図１１】

図11


1101
全体処理状況
処理配列数: 654,321
生成クラスタ数: 56,789
平均クラスタサイズ: 11.5
1102
1103
1104
0%
100%
1106
1105
1121
クラスタアセンブル状況
マッチ数
1109
1107
1108
1112
1111
1110
1122
配列比較状況
長さsの完全一致領域の配列: ATGGC
オーバーラップ長: 123
1113
1123
同時表示
一時停止
1114
1115
1116

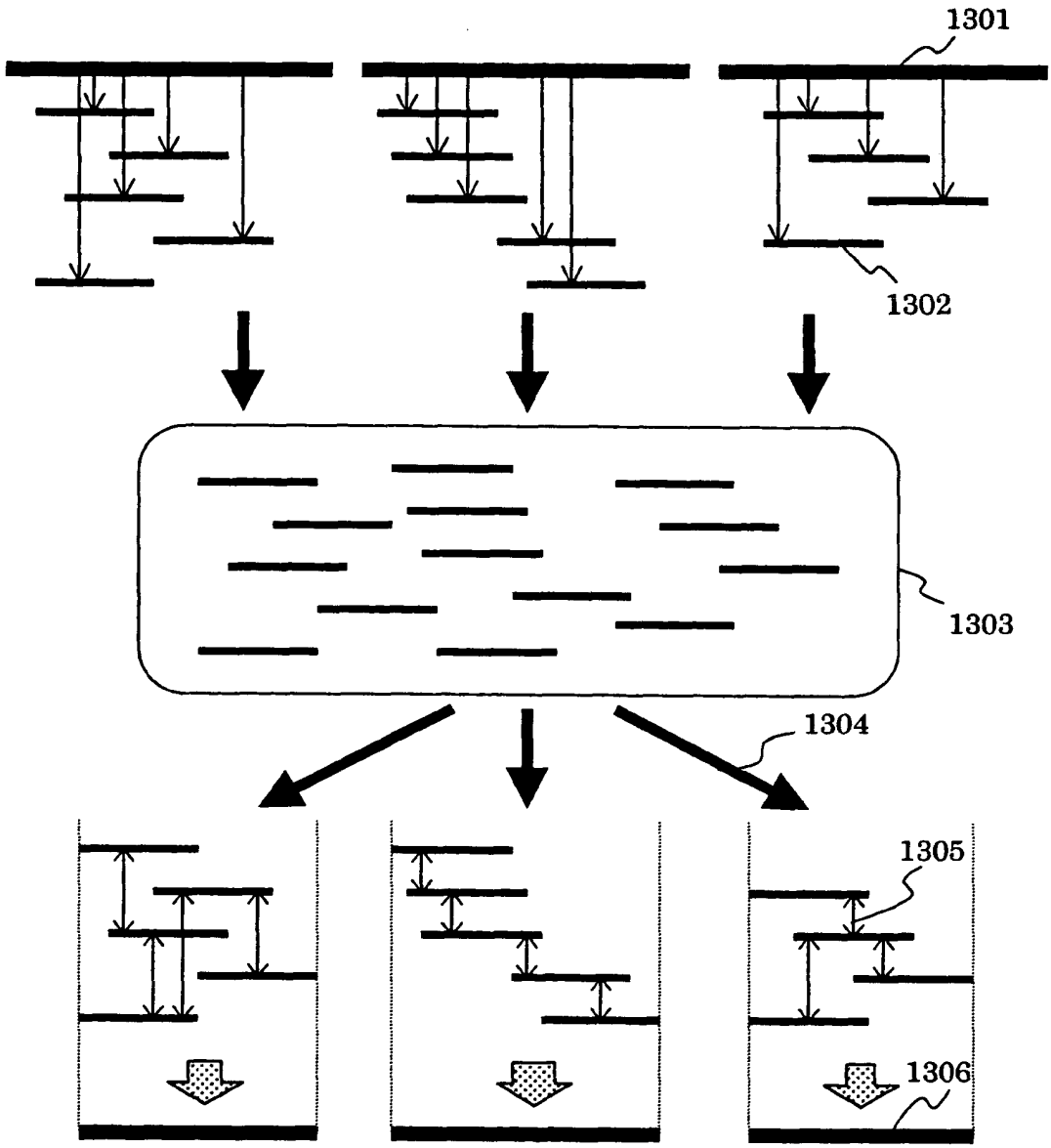
【図１２】

図12


1201
全体の処理結果
入力配列数: 654,321
生成クラスタ数: 56,789
平均クラスタサイズ: 11.5
1222
クラスタアセンブル状況
マッチ数
1205
1204
1223
1208
1207
1206
配列アセンブル状況
1224
オーバーラップ長: 123
1209
長さsの完全一致領域の配列: ATGGC
1210
表示クラスタ選択
1225
入力配列:
1211
表示
1212
クラスタ番号:
表示
1214
1213

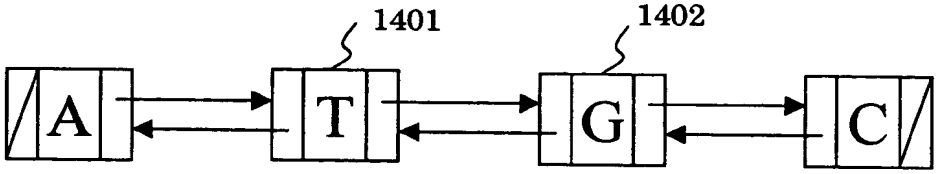
【図１３】

図13



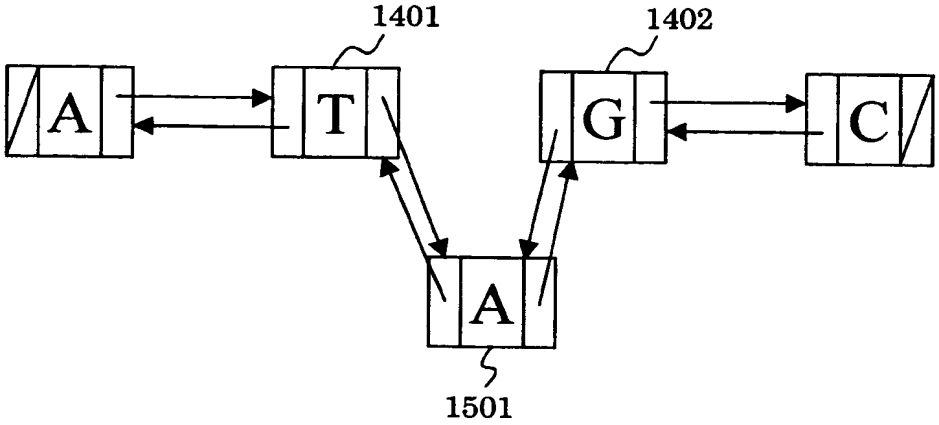
【図１４】

図14



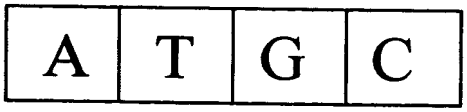
【図１５】

図15



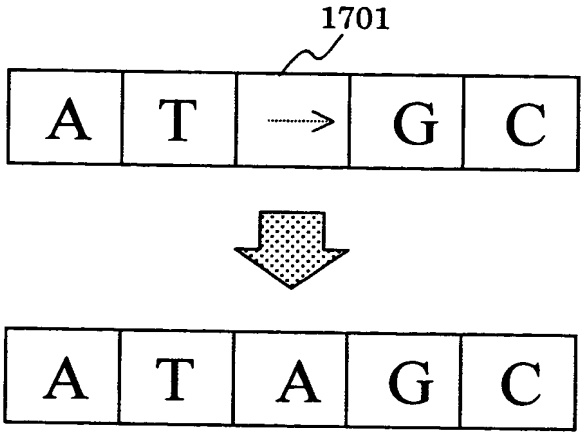
【図１６】

図16



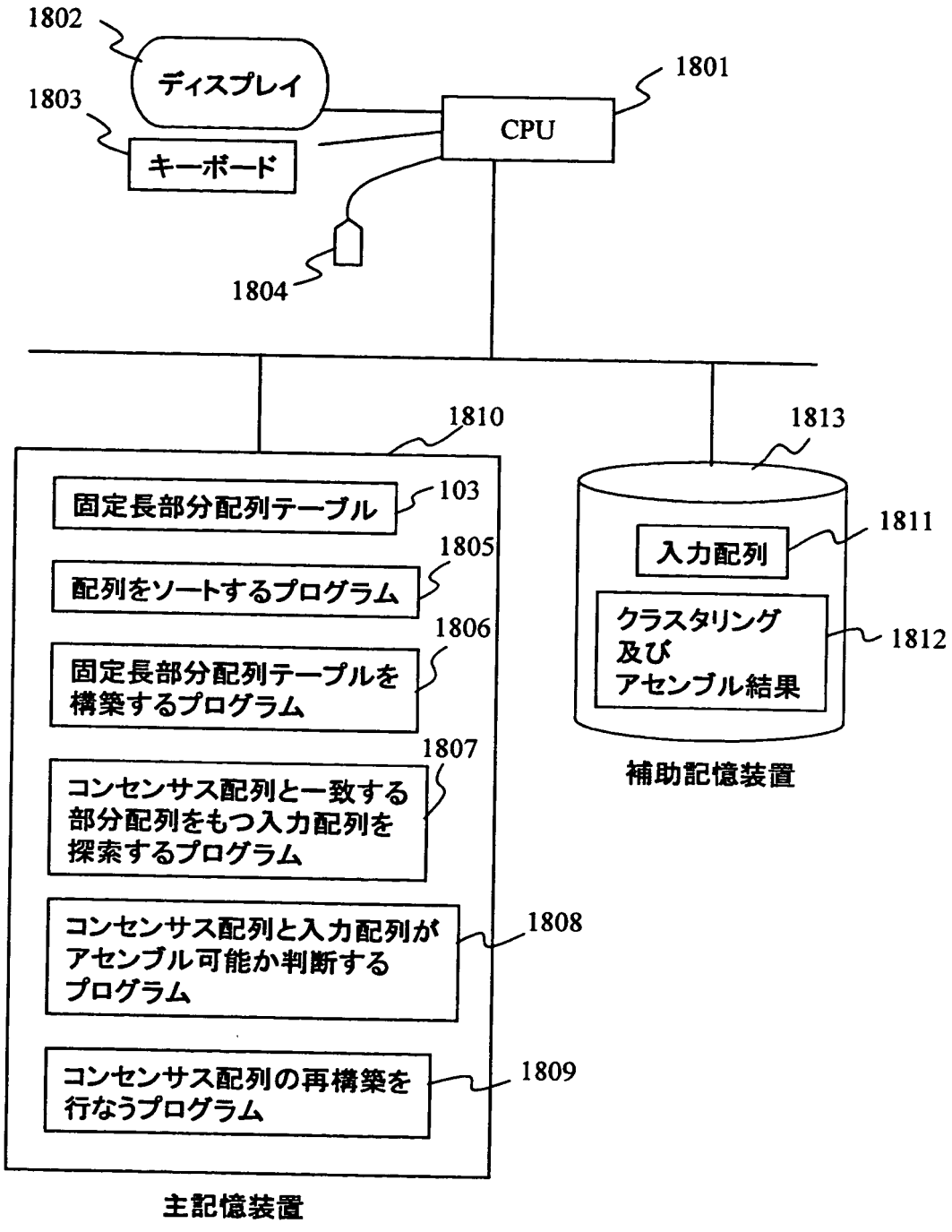
【図１７】

図17



【図１８】

図18


1802
ディスプレイ
1803
キーボード
1801
CPU
1804
1810
1813
固定長部分配列テーブル
103
1805
配列をソートするプログラム
1806
固定長部分配列テーブルを
構築するプログラム
1807
コンセンサス配列と一致する
部分配列をもつ入力配列を
探索するプログラム
1808
コンセンサス配列と入力配列が
アセンブル可能か判断する
プログラム
1809
コンセンサス配列の再構築を
行なうプログラム
主記憶装置
1811
入力配列
クラスタリング
及び
アセンブル結果
1812
補助記憶装置

【書類名】　要約書

【要約】

【課題】核酸塩基配列のクラスタリング及びアセンブルを高速に行う。

【解決手段】個々の入力配列101からその部分配列102を抽出して固定長部分配列テーブル103に記録する。コンセンサス配列104とオーバーラップする配列を探索する際にこの固定長部分配列テーブルを参照し、コンセンサス配列上を走査する固定長ウィンドウ105内の配列に完全一致する部分配列102が存在する場合、配列比較して入力配列全体がアセンブル可能であるかどうか判定する。アセンブル可能である場合に、その配列をクラスタに統合する処理を貪欲法に基づき繰り返し行うことで、クラスタリング及びアセンブルを行う。

【選択図】図１

出 願 人 履 歴 情 報

識別番号　　〔０００００５１０８〕

１．変更年月日　　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所